Universidade do Minho
Escola de Ciências

Alice Maria Salgado Gonçalves

Metodologias estatísticas aplicadas à relação entre Eventos Climáticos Extremos, Saúde e Desigualdades Socioeconómicas na Grande Área Metropolitana do Porto

Outubro de 2012

Metodologias estatísticas aplicadas à relação entre Eventos Climáticos Extremos, Saúde e Desigualdades Socioeconómicas na Grande Área Metropolitana do Porto
Alice Maria Salgado Gonçalves

UMinho | 2012

Universidade do Minho
Escola de Ciências

Alice Maria Salgado Gonçalves

Metodologias estatísticas aplicadas à relação entre Eventos Climáticos Extremos, Saúde e Desigualdades Socioeconómicas na Grande Área Metropolitana do Porto

Tese de Mestrado
Estatística

Trabalho efetuado sob a orientação de

Professora Doutora Susana Faria
Universidade do Minho
Professora Doutora Raquel Menezes
Universidade do Minho
Professora Doutora Ana Monteiro
Universidade do Porto

Outubro de 2012

Este trabalho é financiado por Fundos FEDER através do Programa Operacional Factores de Competitividade – COMPETE e por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do Projeto Ondas com a referência: PTDC/SAU-ESA/73016/2006.

AGRADECIMENTOS

Ao José Pinho pela insistência e conforto para que eu levasse este projecto até ao fim.

Às minhas irmãs Glória (pela colaboração especial), Conceição, Vitória, Brasilina, ao meu irmão Ricardo Paulo e aos meus pais Olívia e João.

À Professora Doutora Susana Faria e à Professora Doutora Raquel Menezes pela sua dedicação e paciência com que me ensinaram as matemáticas e pelo extraordinário empenho para que este trabalho fosse realizado com sucesso.

Agradeço à Professora Doutora Ana Monteiro do Centro de Investigação de Geografia da FLUP pela proposta de trabalhar no tema e por todo o apoio.

À Equipa do Centro de Investigação do Departamento de Geografia da FLUP: o professor Góis da FEUP e os investigadores Vânia, Miguel, Mário, Luís, Sara e Mónica, por todo o apoio e convívio maravilhoso.

A todos muitas graças!!!!!!!

"Obstáculos são as coisas assustadoras que encontra

quando desvia os olhos do seu sonho"

Henry Ford

RESUMO

O presente trabalho intitulado "Metodologias estatísticas aplicadas à relação entre Eventos Climáticos Extremos, Saúde e Desigualdades Socioeconómicas na Grande Área Metropolitana do Porto*"* funda-se no relatório de estágio curricular, sendo parte integrante e conclusiva do curso de Mestrado em Estatística, ministrado pela Escola das Ciências da Universidade do Minho. Este projeto foi desenvolvido no Departamento de Geografia da Faculdade de Letras da Universidade do Porto e está inserido no Projeto Ondas, um projeto financiado por Fundos FEDER e pela Fundação para a Ciência e a Tecnologia.

O presente relatório está estruturado em cinco capítulos. No capítulo inicial faz parte a introdução, no qual é apresentado o tema de estudo e investigação, os objetivos com referência às metodologias estatísticas e ao *software* a utilizar.

O segundo capítulo contextualiza os eventos climáticos extremos, os poluentes ozono (O_3) e as partículas de matéria até 10 micrómetros de diâmetro (PM_{10}), referindo as suas origens e a relação destes fatores com a saúde humana. Refere, ainda, alguns estudos efetuados neste contexto.

O terceiro capítulo está vocacionado para a descrição matemática das várias metodologias e modelos estatísticos utilizados neste estudo, designadamente, os modelos lineares generalizados, os modelos aditivos generalizados e os modelos de regressão de séries temporais para dados de contagem, com distribuição de Poisson.

O capítulo quarto faz uma descrição dos dados e apresentação de resultados. É descrita a proveniência dos dados, é efetuada uma análise descritiva e seguidamente são apresentados e comparados os resultados obtidos através dos diferentes métodos estatísticos.

O quinto capítulo corresponde à conclusão, na qual são efetuadas algumas referências aos capítulos anteriores e às suas principais conclusões, são ainda apresentadas algumas considerações sobre a concretização dos objetivos do projeto de estágio. Neste capítulo são também feitas algumas recomendações ou sugestões para trabalhos futuros.

Palavras-chave: Onda de calor, temperatura aparente, mortalidade, poluição do ar**,** modelos lineares generalizados, modelos aditivos generalizados, séries temporais**.**

ABSTRACT

The present report with the title "Statistical methods applied to the relationship between Extreme Weather Events, Health and Socioeconomic inequalities in the Greater Metropolitan Area of Porto" is based on the training period of two-year-degree of Master in Statistics at School of Sciences, University of Minho. This project was developed at Department of Geography of the Faculty of Arts of the University of Porto and is inserted in Waves Project, a project funded by Funds FEDER and the Foundation for Science and Technology.

The present report is divided into five chapters. The first chapter includes the introduction, which presents the theme of study and the research objectives with reference to the statistical methodologies and software to be used.

The second chapter contextualizes extreme weather events, the pollutants ozone (O_3) and the particular matter (PM_{10}), referring to its origins and the relationship of these factors with human health. It also presents some studies conducted in this context.

The third chapter is devoted to the mathematical description of the ok various methodologies and statistical models used in this study, namely, the generalized linear models, generalized additive models and the of time series regression models for count data, with Poisson distribution.

The fourth chapter gives a description of the data and presents results. The data provenance is explained, a descriptive analysis is performed and the results obtained using different statistical methods are then presented and compared.

The fifth chapter corresponds to the conclusion, in which some references are made to earlier chapters and their main conclusions. Moreover, some comments are made about these objectives which were reached as defined. This chapter also presents some recommendations and suggestions for future works.

Key-words: Heat wave, heat index, mortality, air pollution, generalized linear models, generalized additive models, time series.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

ABREVIATURAS

AIC - *Akaike information criterion*

AR – Autoregressivo

ARMA – Autoregressivo de médias móveis

FAC - Função de autocorrelação

FACP – Função de autocorrelação parcial

FLUP – Faculdade de Letras da Universidade do Porto

GAMP - Grande Área Metropolitana do Porto

HI – *Heat índex*

IC – Intervalo de confiança

IM – Instituto de Meteorologia

INE – Instituto Nacional de Estatística

MA – Médias móveis

MAG – Modelo aditivo generalizado

MLG – Modelo linear generalizado

MMV - Método da máxima verosimilhança

O_3 – Ozono

PAR - Modelo autoregressivo de Poisson

PEWMA[1] – Modelo de Poisson de médias móveis exponencialmente ponderadas

PM_{10} – Partícula de matéria até 10 micrómetros de diâmetro

UM – Universidade do Minho

UP – Universidade do Porto

[1] Do inglês *Poisson exponential weighted moving average model.*

CAPÍTULO I

1 INTRODUÇÃO

1.1 Eventos climáticos extremos e a saúde humana

A Bioclimatologia[2] é a ciência que estuda o efeito do clima nos organismos vivos, sejam plantas, animais ou seres humanos. O ser humano é afetado, de forma positiva e negativa, pelo tempo e pelas condições atmosféricas (Hile, 2009). Embora Hipócrates (377 a. c.) tenha abordado esta questão há mais de 2000 mil anos no tratado *Air, Waters, and Places*, a bioclimatologia como ciência é relativamente recente, tendo um significativo desenvolvimento nos anos 1960 quando aumentou a preocupação com a deterioração do ambiente e os efeitos do clima na saúde humana (Publishing, 2011). Em Portugal, segundo Alcoforado *et al.* (1999), surgiram os primeiros estudos na segunda metade do século XIX. Hile (2009) refere que a Europa já reconhece há algum tempo a importância da Bioclimatologia e como resultado as previsões meteorológicas incluem avisos da possibilidade de riscos para a saúde. Nos Estados Unidos da América, acrescenta Hile, não têm tanto esta preocupação, no entanto, os meteorologistas noticiam com frequência o grau de poluição do ar e os níveis que podem causar alergias, assim como, temperaturas extremas que podem ser perigosas por congelamento ou insolação.

O impacto das alterações climáticas no conforto e na saúde humana é uma preocupação mundial. Os estudos científicos nesta área têm aumentado de forma exponencial, para tentar encontrar relações entre os eventos climáticos extremos e o internamento hospitalar ou a mortalidade, por diferentes doenças, principalmente do foro respiratório, circulatório e neuronal. Citando alguns exemplos, Semenza *et al.* (1999) estudaram o excesso de admissões hospitalares durante a onda de calor de Julho de 1995, em Chicago, enquanto, Braga *et al.*

[2] Ou Biometeorologia, normalmente os autores consideram que são sinónimos (Andrade, 2003)

(2002) analisaram o efeito climático nas doenças respiratórias e cardiovasculares em 12 cidades dos Estados Unidos. Em 2008, Liang *et al.* procuraram efeitos da amplitude diurna na admissão nas urgências por doença crónica pulmunar obstrutiva, em Taiwan. Em 2009, Wang *et al.* investigaram a variação da temperatura e as admissões nas urgências hospitalares por acidente vascular cerebral, no periodo de 1996-2005, em Brismane na Austrália. Na Europa, podemos evidenciar os estudos de Wilkinson *et al.* (2004) que examinaram a mortalidade no inverno na população mais idosa em *Britain*, ou a investigação de Montero *et al*. (2010) sobre a a mortalidade nas vagas de frio em Castile, La Mancha em Espanha. Em Portugal, salientam-se os estudos de Marques e Antunes (2009) que estudaram a perigosidade natural da temperatura do ar em Portugal Continental numa avaliação do risco na mortalidade, Almeida *et al.* (2010) que avaliaram o efeito da temperatuda aparente diária na mortalidade em Lisboa e no Porto e Monteiro *et al.* (2012) que investigaram o excesso de mortalidade e de morbilidade durante o mês de julho de 2006, na Grande Area Metropolitana do Porto.

A área da epidemiologia e mais especificamente a estimativa de risco para a saúde humana associado com a exposição a agentes ambientais, conduziu ao desenvolvimento de vários métodos estatísticos e de *software*. Vários tipos de estudos bioestatísticos foram introduzidos para estimar os efeitos da poluição do ar e dos fatores meteorológicos, na saúde humana. Muitos desses estudos baseiam-se em quatro tipos de metodologias: modelos lineares generalizados, modelos aditivos generalizados, séries temporais e menos frequente mas com relevância a estimação de equações generalizadas.

Conceitualmente, os estudos bioclimáticos coletam dados no tempo e no espaço, de diferentes tipos de exposição dos indivíduos e fatores de confundimento, abrangendo, tipicamente projetos bioestatísticos. Neste trabalho ilustramos a aplicação de modelos estatísticos, replicando importantes estudos no campo de influência dos fatores poluentes e ambientais na mortalidade.

1.2 Objetivos

Este relatório é a consequência do estágio curricular desenvolvido no Departamento de Geografia da Faculdade de Letras da Universidade do Porto (FLUP), sob a orientação da Professora Doutora Susana Faria e da Professora Doutora Raquel Menezes, da Universidade do Minho (UM) e da Professora Doutora Ana Monteiro da Faculdade de Letras da

Universidade do Porto (FLUP), como requisito para obtenção do grau de Mestre em Estatística, na Escola de Ciências da Universidade de Minho.

O estágio está inserido no Projeto Ondas[3] que é um projeto financiado pela FCT – Fundação para a Ciência e Tecnologia, com o título "Riscos para a saúde humana causados pelas ondas de calor e vagas de frio: estudo de caso no Porto", tendo como áreas científicas de interesse a saúde pública, a epidemiologia, os paroxismos climáticos (ondas de calor e vagas de frio) e o planeamento urbano sustentável. O Projeto Ondas tem por objectivo encontrar relações de causalidade existentes entre o agravamento de patologias do foro cardiovascular, respiratório, alergológico, neuronal, etc. e a ocorrência de eventos climáticos extremos, nomeadamente, ondas de calor e vagas de frio, na Grande Área Metropolitana do Porto (GAMP).

O tema deste estudo, intitulado "Metodologias estatísticas aplicadas à relação entre Eventos Climáticos Extremos, Saúde e Desigualdades Socioeconómicas na Grande Área Metropolitana do Porto", foca a compreensão da influência do contexto climático que engloba as condições atmosféricas referentes à temperatura, aparente e aos poluentes, neste caso o ozono (O_3) e as partículas de matéria até 10 micrómetros de diâmetro (PM_{10}), na mortalidade humana, comparando várias metodologias estatísticas. Após realizar uma revisão da literatura mais recente sobre a influência dos fatores ambientais na saúde, neste projeto fazemos uma análise estatística aplicando as metodologias mais utilizadas nesta área e mais adaptadas à natureza dos dados:

- os modelos lineares generalizados (MLG´s),
- os modelos aditivos generalizados (MAG´s) e
- as regressões de séries temporais, para dados com distribuição de Poisson.

No tratamento e análise de dados utilizamos o *software* R.

[3] https://sites.google.com/site/projectondas/informacoes-sobre-o-projecto

CAPÍTULO II

2 ENQUADRAMENTO TEÓRICO

2.1 E quando o clima comporta um risco para a saúde …

A saúde humana é afetada pelos eventos climáticos extremos, como a temperatura chuvas e ventos. Estes acontecimentos têm-se registado por todo o mundo e tendencialmente têm aumentado devido às alterações climáticas. Eventos meteorológicos extremos que têm ocorrido nas últimas décadas, incluindo as ondas de calor de 1995 em Chicago, com centenas de mortes e as ondas de calor do verão de 2003 e 2006 na Europa com milhares de mortes (Skorton, 2011), têm sido responsáveis por um aumento significativo de mortes e doenças. McGlade & Bertollini (2005) falam do exemplo da onda de calor no verão de 2003, em França, que provocou uma *health crisis* com as autoridades de saúde desprevenidas, notando-se a falta de infra-estruturas e recursos humanos. As pessoas, também em geral, foram surpreendidas e não tinham, por exemplo, ar condicionado nas habitações. Esta crise foi agravada pelo facto de muitos idosos viverem sozinhos e não terem capacidade de se protegerem do tempo quente.

Em Portugal, o verão de 2006, referenciado pelo Instituto de Meteorologia (IM) português como os meses de junho, julho e agosto, foi o quinto verão mais quente desde 1931, depois de 2005, 1949, 2004 e 2003. No ano de 2006, no período de 24 de maio a 9 de setembro, verificaram-se 5 ondas de calor, a mais significativa[4] desde 1941, foi registada de 7 a 18 de julho. Tendo ocorrido outra onda de calor de 2 a 11 de agosto, registando-se em alguns dias consecutivos valores da temperatura máxima muito elevados com desvios em relação aos seus valores médios que, em alguns dias, foram superiores a 10°C (Instituto de Meteorologia, 2006).

[4] A onda de calor foi a mais significativa, desde 1941, tanto pela sua extensão espacial que atingiu quase todo o território, como pela sua extensão temporal que permaneceu durante11 dias na região do Alentejo (Instituto de Meteorologia, 2006).

A Organização Mundial da Saúde define uma onda de calor como um episódio com a temperatura relativa acima de 5ºC do valor normal para o período de referência e com uma duração mínima de 4 dias (WHO, 2009). As ondas de calor caracterizam-se pela sua longa duração e intensidade e têm um grande impacto na mortalidade. As ondas de calor aumentam a mortalidade por problemas cardiovasculares e doenças respiratórias. Vários estudos relatam que os dias com altas temperaturas estão associados com o aumento da morbilidade e mortalidade (Braga *et al*., 2002; Díaz *et al*., 2006; Monteiro *et al*., 2012)

Saliente-se que a ocorrência de 3 dias em que a temperatura seja 10 °C acima da média terá certamente mais impacto na saúde que 7 dias com temperatura 5 °C acima da média (Instituto de Meteorologia, 2006).

2.2 A poluição do ar e os riscos para a saúde

Recentes estudos bioestatísticos têm fortalecido a evidência de que a exposição diária a níveis excessivos de poluentes atmosféricos como o ozono (O_3) e as partículas de diâmetro inferior a 10 micrómetros (μm) (PM_{10}), nos grandes centros urbanos, aumentam a taxa de mortalidade e de morbilidade por doenças respiratórias. O ambiente urbano e o efeito de "ilha de calor urbana" fazem aumentar os eventos de temperaturas elevadas e afeta a exposição humana a poluentes (Hoyt *et al*., 2009). O aquecimento do clima e doenças relacionadas com o calor são frequentes nas cidades, porque as cidades aquecem o clima local (Gurjar *et al*., 2008).

O ozono é um composto altamente oxidativo formado na atmosfera inferior a partir de gases por fotoquímica impulsionado pela radiação solar, que pelas suas propriedades químicas altamente reativas, torna-se prejudicial para a saúde. Existem vários níveis estratosféricos de ozono que servem para proteger o planeta das radiações ultravioletas no verão, ao nível terrestre o ozono é produzido pela combinação de vários poluentes que provêm dos automóveis, das lareiras e vários produtos químicos voláteis como solventes, tintas e outros materiais químicos (APA, 2006).

As partículas de diâmetro inferior a 10 micrómetros (μm) constituem um elemento de poluição atmosférica. São partículas inaláveis que podem ser absorvidas para o aparelho respiratório provocando variadas doenças do foro respiratório (Borrego, 1995). Este poluente

pode ser composto por poeira, fuligem, fumo e outros detritos no ar. Fontes de $PM10$ incluem lareiras, incêndios, fumo de tabaco, fábricas e automóveis (APA, 2006).

Em exposições de longo prazo a níveis elevados de ozono, há evidências epidemiológicas e estudos experimentais em animais sobre efeitos nocivos em respostas a inflamações, danos nos pulmões, danos persistentes na estrutura das vias aéreas e alterações do tecido pulmonar no início da vida. Os efeitos nocivos do ozono na saúde incluem 21000 mortes prematuras anuais, associadas aos níveis de ozono quando excedem 70 $\mu g/m^3$ e incluem 14000 admissões hospitalares por causas respiratórias anuais, em 25 países da União Europeia (Amann, 2008). Parece não haver dúvida sobre os efeitos da poluição do ar no aumento da mortalidade. Os acontecimentos em Meuse Valley em Donodora em 1930, na Pensilvânia em 1948 ou o episódio de Dezembro de 1952 em Londres, são bons exemplos (Bates, 1994).

O desenvolvimento de estudos epidemiológicos datam do episódio de Londres, mas foi em 1965 que Holland e Reid (Bates, 1999) criaram um modelo para estudos futuros, quando compararam os efeitos da poluição nos trabalhadores dos serviços postais da região Metropolitana de Londres com outros trabalhadores em cidades menos poluídas, utilizando um modelo *cross-sectional.* Na arquitetura deste estudo, os investigadores asseguraram que o nível socioeconómico destes trabalhadores era o mesmo; não existiam diferenças climáticas nas diferentes regiões; e classificaram os indivíduos em grupos de não-fumadores, ex-fumadores e fumadores. Os resultados evidenciaram um claro decréscimo em todas as categorias entre os trabalhadores de Londres e das províncias. Mas foi nos anos 1970 que se tornou claro que o ozono é um gás intensamente irritante e que as pessoas têm diferentes graus de sensibilidade. A partir desta data realizaram-se diversos estudos em crianças em campos de férias no verão, nos Estados Unidos, que chegaram a duas importantes conclusões: que após a respiração de ozono a capacidade vital torna-se forçada porque os pulmões estão inflamados; o estudo do banco de dados das admissões hospitalares que revelou uma associação significativa entre os níveis de ozono no verão e as admissões hospitalares por doenças respiratórias (Bates, 1999).

A partir dos anos 1990 assistimos a um elevado número de estudos centrados nas partículas urbanas com menos de 10 μm de diâmetro, PM_{10}. Os primeiros dados analisados através de séries temporais, mostraram uma asssociação entre a mortalidade diária, excluindo

acidentes e suicídios, e o nível de PM_{10}. Há um grande concenso de que os níveis de PM_{10} estão associados com resultados menos bons nos testes em crianças, aumento do absentismo escolar, admissões hospitalares por doenças respiratórias, agravação da asma, etc. (Bates, 1999).

Em estudos de curto prazo da morbilidade e da mortalidade, o ozono aparece como tendo um efeito independente de outros poluentes como as PM_{10}. Esta noção de que o ozono atua de forma independente é reforçada por estudos controlados que mostram que tanto no ser humano como em estudos experimentais com animais, o potencial do ozono *per si* causa efeitos adversos na saúde, principalmente nos grupos mais vulneráveis (Amann, 2008). Amann sublinha que os estudos controlados que combinam o ozono e as PM_{10} corroboram este ponto de vista. Em contraste com os estudos com ozono, os que consideram o mecanismo dos efeitos das PM_{10} não estão completamente esclarecidos. Bates (1999) refere que, embora em alguns estudos se tenha notado uma associação quando os níveis de PM_{10} não excedem 150 $\mu g/m^3$ durante um periodo monitorizado a qualquer hora, estes níveis são ainda muito baixos quando comparados com os que muitos trabalhadores em várias áreas são expostos. Também *indoor* os níveis de PM_{10} aumentam consideravelmente quando existe um fumador. Os efeitos nocivos na saúde humana destes poluentes são principalmente doenças respiratórias, que podem provocar morte prematura (APA, 2006).

Reconhecendo a necessidade de reduzir a poluição do ar para níveis que minimizem os efeitos prejudiciais na saúde humana, a diretiva 2008/50/CE do Parlamento Europeu e do Conselho da União Europeia de 21 de maio de 2008 propõe a implementação de uma série de medidas a adotar pelos países membros para melhorar ou manter a qualidade do ar, fixando valores–limite dos poluentes, baseados em conhecimentos científicos, com o objetivo de evitar, prevenir ou reduzir os efeitos nocivos na saúde humana. De acordo com esta diretiva, as partículas em suspensão PM_{10}, no seu limiar de avaliação superior, não devem exceder em média por um período de 24 horas, 70% do valor-limite (35 $\mu g/m^3$, a não exceder mais de 35 vezes por ano civil) e não devem exceder em média anual, 70% do valor-limite (38 $\mu g/m^3$). No seu limiar de avaliação inferior, não devem exceder em média por um período de 24 horas, 50% do valor-limite (25 $\mu g/m^3$, a não exceder mais de 35 vezes por ano civil) e não devem exceder em média anual, 50% do valor-limite (20 $\mu g/m^3$). Em relação ao ozono, o Conselho Europeu recomenda que os valores-alvos de 120 $\mu g/m^3$ não devem exceder mais de 25 dias, em média, por ano civil, tendo como período de referência a média diária por períodos de 8 horas.

Este relatório proporciona um estudo do risco para a saúde, nomeadamente a mortalidade, quando as pessoas são expostas a temperaturas e poluição excessivas.

CAPÍTULO III

3 METODOLOGIAS ESTATÍSTICAS

3.1 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados, designados por MLG´s, foram introduzidos por Nelder & Wedderburn (1972). Estes modelos são uma extensão do modelo linear clássico $Y = X\beta + \varepsilon$, onde *Y* é uma matriz de dimensão $n \times 1$ das observações da variável resposta, *X* é uma matriz de dimensão $n \times (p+1)$ das observações das variáveis explicativas, associada a um vetor $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ de $p+1$ parâmetros e ε é um vetor de erros aleatórios com distribuição que se supõe $N(\mathbf{0}, \sigma^2 \boldsymbol{I})$. No entanto, McCullagh & Nelder (1989) salientam que vários modelos não lineares ou não normais, mas com uma estrutura de regressão linear, já tinham sido desenvolvidos para fazer face a situações que não eram adequadamente explicadas pelo modelo linear normal, como o modelo complementar *log-log* para ensaios de diluição (Fisher, 1922), o modelo *probit* (Bliss, 1935), o modelo *logit* para proporções (Berkson, 1944; Dyke and Patterson, 1952), os modelos *log-lineares* para dados de contagem (Birch, 1963) ou os modelos de análise de sobrevivência (Feigl & Zelen, 1965; Zippin & Armitage, 1966; Glasser, 1967).

A extensão mencionada é feita em duas direções (Turkman & Silva, 2000):

i) – A distribuição da variável resposta não tem que ser normal podendo ser qualquer distribuição da família exponencial;

ii) – Embora se mantenha a estrutura de linearidade, a função que relaciona o valor esperado da variável resposta e o valor das variáveis explicativas pode ser qualquer função diferenciável.

Os MLG´s, pressupõem que a distribuição da variável resposta Y pertence à família exponencial cuja função densidade de probabilidade, ou função massa de probabilidade, pode ser expressa da seguinte forma

$$f(y|\theta,\emptyset) = \exp[\frac{y\theta - b(\theta)}{a(\theta)} + c(y,\emptyset)]. \quad (3.1)$$

onde o θ é o parâmetro canónico (natural)[5] que representa a localização e o $\emptyset$ é designado por parâmetro de dispersão e representa "a *escala exigida para produzir erros padrão seguindo uma distribuição pertencente à família exponencial de distribuições* …" (Hardin & Hilbe, 2007). As funções $a(.), b(.)$ e $c(.)$ são funções reais conhecidas e específicas para cada distribuição que definem os vários membros da família.

A distribuição normal ou Gaussiana, a Poisson, a binomial, a exponencial, a gama, a Gaussiana inversa, a geométrica e a binomial negativa, são alguns exemplos de distribuições que pertencem à família em estudo.

Se Y é uma variável aleatória com uma distribuição que pertence à família exponencial, então, a média e a variância, são dadas respetivamente, por $E[Y] = \mu = b^{'}(\theta)$ e $Var\,[Y] = b^{''}(\theta)a(\emptyset)$, onde $b^{'}(\theta)$ e $b^{''}(\theta)$ são a primeira e a segunda derivadas de $b(\theta)$, respetivamente.

A média é uma função de θ, enquanto que a variância é o produto de duas funções: de localização e de escala. O termo $b^{''}(\theta)$ designa-se por função de variância que se costuma representar por $V(\mu)$ e descreve de que forma a variância se relaciona com a média.

A distribuição Gaussiana, é um caso particular, em que $b^{''}(\theta) = 1$, e portanto, a variância é independente da média.

Em muitas situações, a função $a(\emptyset)$ é da forma $\frac{\emptyset}{w}$, onde w é um peso constante conhecido.

[5] Também é conhecido como *parâmetro de localização*.

3.1.1 Componentes de um modelo linear generalizado

Os MLG´s são caracterizados por três componentes distintas (McCullagh & Nelder, 1989):

i) – Componente aleatória que identifica a variável resposta Y e especifica uma distribuição Y pertencente à família exponencial;

ii) – Componente sistemática que especifica as variáveis explicativas do modelo e considera uma combinação linear dessas variáveis;

iii) – Função de ligação que estabelece a ligação entre a componente aleatória e a sistemática.

Componente aleatória

Dado o vetor de covariáveis $X = (x_1, \ldots, x_p)$ a componente aleatória de um MLG refere que a distribuição de Y condicionada por X pertence à família exponencial com

$E(Y_i|X_i) = \mu_i = b'(\theta_i)$ para $i = 1, \ldots, n$.

Componente sistemática

Num MLG o valor esperado μ_i e o preditor linear $\eta_i = \boldsymbol{X}_i^T\boldsymbol{\beta}$ estão relacionados pela equação $\mu_i = h(\eta_i) = h(\boldsymbol{X}_i^T\boldsymbol{\beta})$, com $\eta_i = g(\mu_i)$, em que h é uma função monótona e diferenciável, $g = h^{-1}$ é a função de ligação, β é um vetor de parâmetros, $\boldsymbol{X}_i$ é o vetor de covariáveis.

A função de ligação

A função de ligação, $g = h^{-1}$ é uma função monótona e diferenciável que relaciona o valor esperado da variável resposta Y, $E[Y] = \mu$, com o predictor linear $\eta = g(\mu)$.

Os casos em que o preditor linear coincide com o parâmetro canónico θ, ou seja, $\theta_i = \eta_i$, implica manifestamente que $\theta_i = \boldsymbol{X}^T\boldsymbol{\beta}$, a função de ligação chama-se função de ligação canónica (Turkman & Silva, 2000). As funções de ligação canónica mais utilizadas, nos MLG´s, são descritas na Tabela 1.

Tabela 1 Funções de ligação canónica para os MLG´s

Família	**Ligação**	**Função variância**
Normal	$\eta = \mu$	1
Poisson	$\eta = \log \mu$	μ
Binomial	$\eta = \log(\mu/(1-\mu))$	$\mu(1-\mu)$
Gama	$\eta = \mu^{-1}$	μ^2
Gaussiana Inversa	$\eta = \mu^{-2}$	μ^3

Fonte: Faraway (2006, Table 6.1, p. 128)

3.1.2 Estimação dos parâmetros

Estimação do vetor de parâmetros $\boldsymbol{\beta}$

Vários métodos podem ser aplicados para estimar o vetor dos parâmetros β do MLG, como o método bayesiano ou a estimação-M, no entanto, o método da máxima verosimilhança (MMV) é o método mais frequente. A estimação dos parâmetros de regressão, os testes de hipóteses sobre os parâmetros do modelo e a qualidade de ajustamento do modelo são, em geral, baseados na verosimilhança. Este método, fornece estimadores consistentes, assintoticamente eficientes e com distribuição assintoticamente normal (Turkman & Silva, 2000).

Maximizar a função de verosimilhança é equivalente a maximizar a *log*-verosimilhança. O logaritmo da função de verosimilhança, como função de β, é dado por

$$ln\, L(\beta) = \sum_{i=1}^{n} \left\{ \frac{w_i\,(y_i\theta_i - b(\theta_i)}{\emptyset} + c(y_i\,, \emptyset, w_i) \right\} = \sum_{i=1}^{n} l_i\,(\beta), \qquad (3.2)$$

em que $l_i\ (\beta) = \frac{w_i\,(y_i\theta_i - b(\theta_i)}{\emptyset} + c(y_i\,, \emptyset, w_i)$ é a contribuição de cada observação y_i para a verosimilhança. Os estimadores de máxima verosimilhança para β são calculados como solução das equações de verosimilhança

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \beta_j} = 0, \text{ onde } j = 1, \ldots, p \text{ e}$$

$$\frac{\partial l_i(\beta)}{\partial \beta_j} = \frac{\partial l_i(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i(\mu_i)}{\partial \mu_i} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} \frac{\partial \eta_i(\beta)}{\partial \beta_j},$$

sendo que

$$\begin{cases} \frac{\partial l_i(\theta_i)}{\partial \theta_i} = \frac{w_i \left(y_i - b'(\theta_i)\right)}{\emptyset} = \frac{w_i (y_i - \mu_i)}{\emptyset} \\ \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{w_i \, var(Y_i)}{\emptyset} \\ \frac{\partial \eta_i(\beta)}{\partial \beta_j} = X_{ij} \end{cases}$$

Desta forma, temos

$$\frac{\partial l_i(\beta)}{\partial \beta_j} = \frac{w_i (y_i - \mu_i)}{\emptyset} \frac{\emptyset}{w_i \, var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \tag{3.3}$$

e as equações de verosimilhança para β são obtidas por

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) z_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, j = 1, \ldots, p. \tag{3.4}$$

As equações de verosimilhança são, de um modo geral, não lineares o que torna ncessário a utilização de métodos iterativos para a sua resolução. Na resolução numérica destas equações de verosimilhança pode-se utiliza o método de *scores* de Fisher. A função *score* é dada por

$$s(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} s_i(\beta), \tag{3.5}$$

em que $s_i(\beta) = \frac{\partial l_i(\beta)}{\partial \beta_j}$.

Portanto, o elemento genérico de ordem *j* da função *score* é

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}. \tag{3.6}$$

A matriz de covariância da função *score* é a matriz de informação de Fisher $I(\beta)$ dada por

$$I(\beta) = E\left[-\frac{\partial S(\beta)}{\partial \beta}\right]. \tag{3.7}$$

Para obter $I(\beta)$ é necessário considerar o valor esperado das segundas derivadas de $l_i(\beta)$,

$$-E\left(\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k}\right) = \frac{x_{ij}x_{ik}}{Var(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{3.8}$$

Na forma matricial $I(\beta) = X^T W X$, em que W é a matriz diagonal de ordem *n* e o *i*-ésimo elemento é dado por

$$W_i = \frac{(\frac{\partial \mu_i}{\partial \eta_i})^2}{var\,(Y_i)} = \frac{W_i(\frac{\partial \mu_i}{\partial \eta_i})^2}{\emptyset\, V\,(\mu_i)}. \tag{3.9}$$

A estimativa de máxima verosimilhança de β na (k+1)-ésima iteração é dada por

$$\hat{\beta}^{(k+1)} = (X^T W^{(k)} X)^{-1} X^T W^{(k)} u^{(k)} \tag{3.10}$$

em que $u^{(k)}$ é um vetor com elemento genérico

$$u_i^{(k)} = \eta_i^{(k)} + (y_i - \mu_i^{(k)})\,\frac{\partial \eta_i{}^{(k)}}{\partial \mu_i^{(k)}} \tag{3.11}$$

O elemento genérico de W contém ∅, no entanto, este parâmetro não é utilizado para o cálculo de $\hat{\beta}^{(k+1)}$ o que permite considerar $\emptyset = 1$, quando se está a calcular as estimativas de β. Assim sendo, é irrelevante, o conhecimento ou não do parâmetro de dispersão, para a estimativa de $\hat{\beta}$.

Estimação do parâmetro de dispersão

O parâmetro de dispersão ∅, pode ser estimado usando o método da máxima verosimilhança ou o método dos momentos. No entanto, existe um método mais simples, baseado na distribuição de amostragem, para grandes valores de dimensão da amostra *n*, da estatística de *Pearson* generalizada, que dá geralmente, bons resultados (Turkman & Silva, 2000). O parâmetro pode ser estimado por

$$\widehat{\emptyset} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{w_i(y_i - \widehat{\mu}_i)^2}{V(\mu_i)} \tag{3.12}$$

sendo $\widehat{\emptyset}$ um estimador de $\emptyset$. Para grandes valores de *n* tem uma distribuição aproximada de um $\mathcal{X}^2$ com *n-p* graus de liberdade.

3.1.3 Modelo de regressão de Poisson

Nos modelos analisados neste estudo, a variável resposta Y são dados de contagem, ou seja, número de óbitos. Quando temos uma variável resposta Y representando dados de contagem, números positivos e inteiros, podemos utilizar um modelo de regressão de Poisson para explicar a relação da variável resposta em função das variáveis explicativas. Os dados de contagem podem seguir uma distribuição de Poisson, ou menos frequente, a binomial negativa. Se considerarmos que as variáveis Y_i são independentes e que seguem por uma distribuição de Poisson de valor médio μ_i e em que $ln(\mu_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}$ com $i = 1, \dots, n$, então um modelo de regressão de Poissson ou modelo *log*-linear, dado por

$$f(y_i|\boldsymbol{X}_i) = e^{-\mu_i}\frac{\mu_i{}^{y_i}}{y_i!} = exp\left\{-e^{\boldsymbol{X}_i^T\boldsymbol{\beta}} + y_i\boldsymbol{X}_i^T\boldsymbol{\beta} - lny_i!\right\}, y_i = 0, 1, \dots, \quad (3.13)$$

Para o modelo de Poisson a função de ligação, é geralmente, a função logarítmica.

No entanto, quando estamos a manusear dados de contagem, o modelo de regressão de Poisson clássico é muitas vezes limitado, porque os dados empíricos apresentam, tipicamente, sobredispersão.

3.1.4 O problema da sobredispersão

A distribuição de Poisson assume, teoricamente, que a média e a variância são iguais, $E(Y) = Var(Y) = \mu$, mas em situações reais nem sempre se verifica essa igualdade. Em dados de contagem, às vezes verifica-se que $Var(Y) > E(Y)$, ou seja, existe sobredispersão no modelo e neste caso, é necessário alterar a variância introduzindo um parâmetro de sobredispersão. A sobredispersão pode verificar-se por várias razões. McCulagh & Nelder (1989) explicam que a sobredispersão pode acontecer porque há um efeito aleatório de cada indivíduo, ou há uma tendência para as observações pertencerem a grupos. Adicionalmente, este fenómeno pode verificar-se quando os dados apresentam correlação temporal (Zeger, 1988; Myers *et al*, 2010).

A sobredispersão pode ser considerada um problema, porque pode inflacionar ou subestimar os erros padrão das estimativas, o que pode levar a que consideremos uma variável preditora como significativa, quando na realidade o não é (Hilbe, 2008). Esta questão pode ser contornada utilizando a função de *quasi*-verosimilhança (McCulagh & Nelder , 1989)

3.1.5 A função de *quasi*-verosimilhança

O termo função de *quasi*-verosimilhança foi introduzido por Wedderburn (1974). A utilização da função de *quasi*-verosimilhança requer a introdução de um parâmetro de sobredispersão $\emptyset > 1$ desconhecido, ou seja, assumir um modelo com

$$Var(Y_i) = \emptyset V(\mu_i). \tag{3.14}$$

No entanto, um modelo com estas características, já não pode ser escrito na forma da família exponencial. Neste caso, o modelo deixa de ser especificado uma vez que apresentando esses valores de média e variância, não existe uma distribuição conhecida dentro da família exponencial e o modelo fica apenas identificado pelo valor esperado e pela variância, não sendo possível recorrer à função de verosimilhança para fazer inferências. Para definirmos a função de verosimilhança temos de especificar a distribuição das observações, mas a função de *quasi*-verosimilhança, somente necessita da especificação da relação existente entre o valor esperado e a variância das observações (Wedderburn, 1974; McCulagh & Nelder , 1989).

Considerando o caso em que Y_i tem valor médio μ_i e variância $\emptyset V(\mu_i)$, assumindo que os valores de Y_i são independentes. Considerando ainda a variável

$$U_i = \frac{Y_i - \mu_i}{\emptyset V(\mu_i)} \tag{3.15}$$

que verifica

$$E[U_i] = 0$$

$$var[U_i] = \frac{1}{\emptyset V(\mu_i)}$$

$$-E\left[\frac{\partial U_i}{\partial \mu_i}\right] = -E\left[\frac{-\emptyset V(\mu_i) - (Y_i - \mu_i)\emptyset V'(\mu_i)}{[\emptyset V(\mu_i)]^2}\right] = \frac{1}{\emptyset V(\mu_i)} \tag{3.16}$$

Tem-se que U_i comporta-se como uma função *score* (ou seja a derivada da função *log*-verosimilhança), como tal, esperamos que o integral de U_i se comporte como uma função *log*-verosimilhança. Isto sugere que a função de *quasi*-verosimilhan*ça* (ou *quasi-log*-verosimilhança) pode ser definida da seguinte forma:

$$Q_i(\mu_i y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} \tag{3.17}$$

No caso em que temos n observações de variáveis aleatórias independentes a função *quasi*-verosimilhança, é o somatório de cada contribuição individual

$$Q(\mu_i y_i) = \sum_{i=1}^{n} Q_i(\mu_i y_i) \tag{3.18}$$

Num modelo de regressão de *quasi*-Poisson o parâmetro $\emptyset$ é estimado através dos dados e não $\emptyset = 1$. Utilizando esta estratégia, a estimação dos parâmetros não é afetada pelo valor de $\emptyset$, ou seja, os parâmetros estimados são os mesmos que no modelo Poisson clássico mas a inferência é ajustada tendo em conta a sobredispersão (Zeileis *et al*., 2008) .

3.1.6 O modelo de regressão binomial negativo

Outra forma de resolver o problema da sobredispersão dos dados, como já foi referido, é utilizarmos um MLG binomial negativo. A distribuição binomial negativa pode ser entendida como uma mistura das distribuições Gama e Poisson (Zeileis *et al.,* 2008), com função densidade probabilidade dada por

$$f(Y_i|X_i) = \frac{\Gamma_{(y+\theta)}}{\Gamma_{(\theta)}.y_i!} . \frac{\mu^{y_i}.\theta^{\theta}}{(\mu+\theta)^{y_i+\theta}} , \tag{3.19}$$

em que μ é a média, θ é o parâmetro de forma e $\Gamma_{(.)}$ é a função gama (Myers *et al.*, 2010).

3.2 Modelos Aditivos Generalizados

No modelo linear generalizado o preditor linear é uma função linear de cada uma das variáveis explicativas $X_1, \ldots, X_p$. O modelo aditivo generalizado (MAG) é uma extensão do modelo linear generalizado em que o preditor linear é substituído pela soma de funções de suavização não paramétricas que sumariam a associação entre a variável resposta e as

variáveis explicativas (Hastie & Tibshirani, 1986, 1990). No entanto, os modelos semiparamétricos cujo preditor combina formas paramétricas de algumas das r variáveis preditoras com termos não paramétricos das outras (p - r) variáveis são também MAG's. De uma forma geral, este modelo apresenta a seguinte estrutura

$$\eta = g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_r X_r + f(X_{r+1}) + \cdots + f(X_d) \quad (3.20)$$

Portanto, os MAG´s permitem especificar a relação entre a variável Y e as covariáveis, de uma forma bastante flexível, e, especificar um modelo em termos de funções suaves, sendo possível evitar modelos complexos (Wood, 2006). Os gráficos das curvas de suavização proporcionam uma visão ótima da relação entre a variável Y e as covariáveis, apresentando as possíveis não linearidades nas relações estudadas.

Os MAG´s são baseados em funções não paramétricas, chamadas curvas de suavização, que são estimadas através de curvas de suavização sobre a variável que se deseja controlar. Estes modelos são uma alternativa, quando temos dados que apresentam relações não-lineares (Hastie & Tibshirani, 1990). A grande vantagem da utilização destes modelos aditivos é que podemos utilizar várias transformações simultaneamente e sem pressupostos paramétricos em relação à sua forma (Faraway, 2006).

Num modelo clássico de regressão, com *n* observações da variável aleatória Y , fazemos corresponder os pontos das observações $y_1, y_2, \dots, y_n,$ aos pontos das observações de um preditor X, $x_1, x_2, \dots, x_n$. O algoritmo da regressão produz uma decomposição sob a forma $y_i = \hat{y}_i + \varepsilon_i$, em que o valor ajustado $\hat{y}_i$ é uma estimação da dependência sistemática de y_i em x_i. Se aplicarmos um método de suavização a esta dependência, o resultado do desenho dos pontos, corresponde a um diagrama de dispersão com uma linha de suavização.

Existem vários métodos de curvas de suavização desenvolvidos. Em qualquer um dos métodos a suavização é obtida ajustando-se uma curva aos dados de tal forma, que a cada ponto, a curva dependa somente das observações naquele ponto (x_0) e em uma vizinhança. Buja *et al*.(1989), referem alguns suavizadores lineares, como por exemplo, a média móvel, *locally weighted running line (loess)*, o método *kernel*, os *smoothing splines*. Também existem métodos de suavização não paramétricos, como a mediana móvel, em que não é possível desenhar matriz dos suavizadores (Buja, Hastie, & Tibshirani, 1989) , ou ainda, os

lowess (Cleveland, 1979). Alguns dos métodos, acabados de referir, estão representados na Figura 1, para dados do número de óbitos em função dos dias transcorridos.

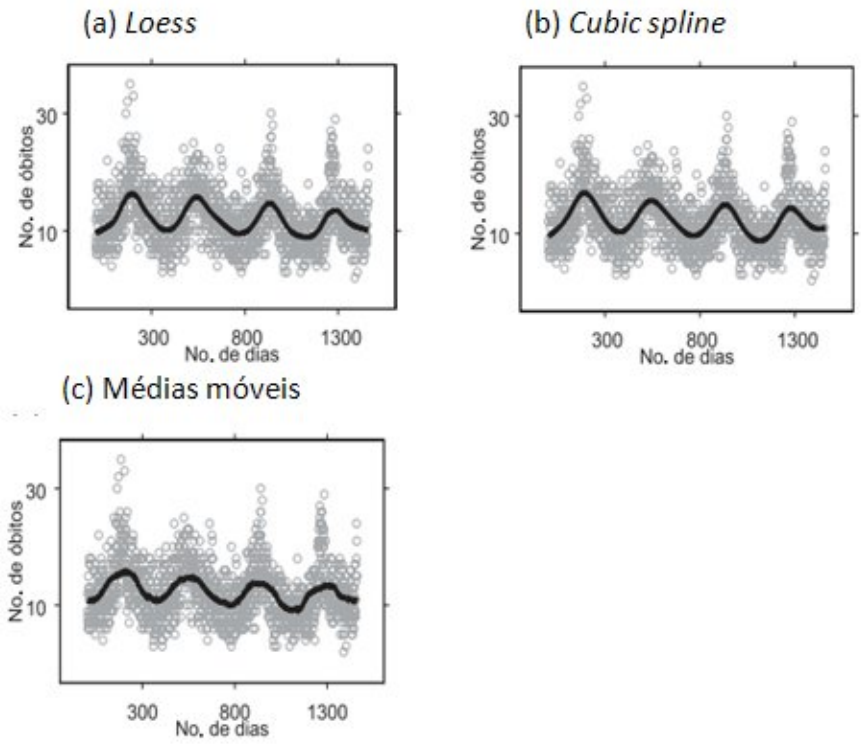

Figura 1 Curvas de suavização do número de óbitos por causas respiratórias em idosos em função do número de dias

3.2.1 Média móvel

O método da média móvel produz um modelo para x_i usando a média dos pontos da vizinhança N_i , que se encontram em torno de x_i. Apenas podemos considerar vizinhos simétricos. O tamanho da vizinhança N_i a considerar, designa-se *span* ou janela e é identificado por w, que é a proporção do total de pontos contido em cada vizinhança. Considere $[x]$ a parte inteira de x e assuma $[w_n]$ é impar, então a janela w poderá conter $[w_n]$ pontos, com $([w_n]$ -1)/2 pontos à direita e à esquerda de x_i e contendo também o x_i (Hastie & Tibshirani, 1986).

Assumindo que os pares de dados são ordenados por ordem crescente, a equação para o cálculo da vizinhança de x_i

$$N_i = \left\{\max\left(i - \frac{[w_n]-1}{2}, 1\right), \ldots, i-1, i, i+1, \ldots, min\left(i + \frac{[w_n]-1}{2}, n\right)\right\}. \qquad (3.21)$$

A janela ($span$) w controla o grau de suavização do resultado estimado, quanto maior mais suave é a função.

3.2.1 O método *Kernel*

O suavizador de *Kernel* é um método não paramétrico para estimação de curvas de densidades onde cada observação é ponderada pela distância em relação a um valor central. O objetivo é centrar cada observação x_i onde se queira estimar a densidade, uma janela que define a vizinhança de x_i e os pontos que pertencem à estimação.

O método de *Kernel* estima uma função "*g*" de vizinhança de cada ponto de interesse $x = x_0$, considerando a média ponderada das observações que estão na vizinhança do ponto de interesse, x_0. Para estimar uma curva de regressão por este método, em primeiro lugar, devemos escolher o tamanho da vizinhança N_i do ponto x_0 e em segundo lugar, escolhemos uma função *k* (denominada por núcleo de *Kernel*) que pondera o conjunto dos pontos vizinhos a x_0. A vizinhança N_i é denominada parâmetro de suavização.

3.2.2 Locally weighted running-line (loess)

O suavizador linear *loess (locally weighted running-line)* constrói uma linha baseada nos mínimos quadrados dos pontos da vizinhança simétrica N_i de cada x_i. O suavizador *loess* é considerado uma generalização simples do suavizador da média móvel e utilizando os mínimos quadrados em alternativa à média em cada vizinhança.

Esta técnica tende a gerar curvas muito irregulares porque os pontos em uma dada vizinhança têm pesos iguais (não nulos) enquanto pontos fora desta região têm peso zero.

3.2.3 Spline cúbico

O *spline* cúbico é das técnicas mais utilizadas em estudos epidemiológicos que englobam os efeitos da temperatura e da poluição na saúde humana (Hu, *et al.*, 2008; Iñiguez, *et a.l*, 2010; Díaz *et al.*, 2006; Almeida *et al.*, 2010).

Um *spline* cúbico é uma curva construída a partir de secções de polinômios cúbicos unidos, para que a curva seja contínua até à segunda derivada. A segunda derivada da função garante que os polinóminos se unam de forma suave.

O fundamento dos *splines* cúbicos é derivar um polinômio de terceiro grau, para cada uma das secções entre dois nós, que são o ponto onde os dois *splines* se encontram, (Hastie & Tibshirani, 1986:1990; Chapra & Canale , 1987; Wood, 2006).

A Figura 2 (Wood, 2006) mostra um *spline* que é constituído por 7 secções de polinómios cúbicos. Os oito pontos (◦) são os nós do *spline*. As linhas retas a tracejado mostram a inclinação do *spline* nos nós. As linhas curvas contínuas são funções quadráticas que correspondem às primeiras e segundas derivadas nos nós: estas linhas ilustram a continuidade da primeira e segunda derivadas dos nós. O *spline* da Figura 2 tem segunda derivada igual a zero no último nó, é um *spline* natural.

Este *spline* consiste em dividir os valores da covariável em intervalos predefinidos e ajustar um polinómio para cada uma deles, considerando uma proporção do conjunto dos valores da covariável para a construção da função de suavização.

O *spline* cúbico é estimado a partir de todas as funções $f(x)$ com segunda derivada que minimizam a soma dos quadrados dos resíduos penalizados, através da equação

$$\sum_1^n (y_i - f(x_i))^2 + \lambda \int_{-\infty}^{+\infty} [f''(x)]^2 \, dx \qquad (3.22)$$

onde λ é o parâmetro de suavização, e o integral $\int_a^b [f''(x)]^2 \, dx$, é constituído por um intervalo [a, b] que contém os valores arbitrários da variável de suavização *x* (Hastie & Tibshirani, 1990; De Boor, 1978). Um método de suavização assume que as observações são independentes e identicamente distribuídas, e em particular que não há correlação temporal (Maindonald J. , 2010).

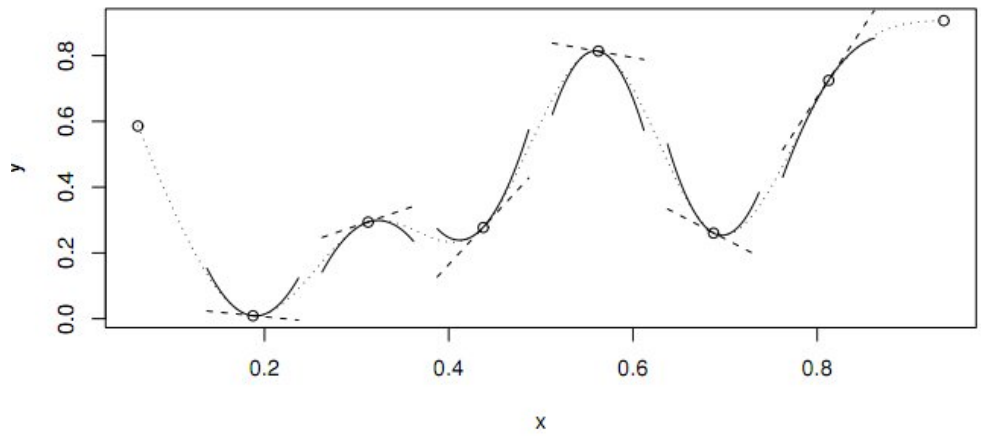

Figura 2 *Spline* cúbico

Fonte: Wood, 2006, Fig. 3.3, p.122)

Funções de suavização com poucos graus de liberdade dificultam a captação dos picos de aumento ou diminuição e são mais enviesados nestas áreas. Com maior número de graus de liberdades e consequentemente mais flexibilidade, podemos reduzir o enviesamento nestas áreas (Peng & Dominici, 2011).

3.3 Séries temporais

Uma série temporal é uma realização de um processo estocástico, podendo ser definida como um conjunto de dados recolhidos sequencialmente ao longo do tempo, não sendo necessariamente igualmente espaçados. Na maior parte dos procedimentos estatísticos supomos que as observações são de uma amostra aleatória, e portanto, são independentes e identicamente distribuídas (*i.i.d.*). Uma das características mais importantes das séries temporais, é que, os dados apresentam estrutura de correlação, ou seja, dependência das observações entre os instantes de tempo. Numa série temporal temos um conjunto de observações $\{y_t: t = 1, \dots, n\}$ em que t indica o tempo em que o dado y_t foi observado (Diggle, 1990) . Uma série temporal pode ser discreta para y_t com $t = 1, \dots, n$ ou contínua para y_t com $t \in \mathbb{R}$.

Enquanto nos modelos lineares generalizados, por exemplo, a ordem das observações é irrelevante, na análise de uma série temporal a ordem dos dados é fundamental. Importa

também referir, que para esta metodologia a variável tempo pode ser substituída por outra variável como o espaço, a profundidade, ou outro conjunto qualquer.

A análise de séries temporais tem como objetivos descrever, explicar, prever e controlar fatores (Chatfield, 1996), torna-se, então, importante descrever a série, ou seja, verificar as suas características. Quando temos duas ou mais variáveis, podemos estar interessados em verificar se a variação de uma série pode explicar a variação de outra série. Se a informação do passado ou do presente de uma variável de interesse contém informação para o desenvolvimento futuro dessa variável, é plausível utilizar para previsão alguma função dos dados recolhidos. Podemos ainda controlar fatores, quando geramos séries temporais para medir a qualidade longitudinal de um fenómeno.

Numa análise de séries temporais pretendemos encontrar o modelo que melhor descreva o comportamento do fenómeno, ou seja, determinar modelos matemáticos que expliquem a variabilidade dos dados, através de funções determinísticas do tempo e das observações passadas, capturando as suas características mais relevantes e ainda as suas possíveis relações com outras séries (Shumway & Stoffer, 2011). Para podermos fazer inferência sobre as séries, antes é necessário criar um modelo de probabilidade hipotética para representar os dados, definindo a família mais adequada e a partir deste ponto ajustar um modelo que melhor faça entender o mecanismo dinâmico dos dados temporais e estimar os parâmetros. Depois de encontrado um modelo satisfatório podemos utilizá-lo de variadas formas com dinâmica e criatividade, dependendo do campo particular de aplicação (Brockwell & Richard, 2002).

Quando assumimos um modelo univariado (Johnston & DiNard, 2007), o comportamento de y_t é apenas explicado pelos seus valores passados e pelos valores passados e presentes de um erro estocástico, evoluindo segundo um processo de ruído branco, ε_t, como mostra a equação

$$Z_t = f(z_{t-1}, z_{t-2}, \dots, \varepsilon_t, \varepsilon_{t-1}, \dots), \text{ com } \varepsilon_t \sim i.i.d.(0, \sigma_\varepsilon^2), \tag{3.23}$$

em que σ_ε^2 é a variância dos erros aleatórios.

3.3.1 Estacionariedade

Matematicamente, diz-se que uma série temporal é estacionária se a distribuição conjunta de n observações, $\{z_{t+1}, z_{t+2}, \dots, z_{t+n}\}$, considerando a realização da série temporal nos tempos $t+1, t+2, \dots, t+n$, permanecer a mesma para um outro conjunto de n observações desfasado k unidades de tempo.

Uma série temporal diz-se francamente estacionária, na prática trata-se de uma assunção mais fácil de garantir, se apresenta uma média e variância constantes no tempo e a correlação entre as observações, para os diferentes pontos no tempo, é unicamente dependente dos desfasamentos. Por outras palavras podemos dizer que uma série z_t é um processo estritamente estacionário, se $\mathrm{E}[z_t] = \mu$ não depende do tempo t a autocovariância $\mathrm{Cov}[z_{t+k}, z_t]$ depende unicamente de k para todo o t (Brockwell & Richard, 2002; Bisgaard & Kulahci, 2011).

A aplicação de modelos autoregressivos (AR) e de média móvel (MA), que são descritos na próxima secção, requer estacionariedade.

Para dados não estacionários existem vários métodos de transformação para que se tornem estacionários. Por exemplo, se a variância não é constante podemos extrair o logaritmo ou uma potência da série; se existir tendência pode-se ajustar o desvio por uma curva, subtraindo o valor da curva à série; ou aplicar métodos de diferenciação, que recorrem ao operador desfasamento. Tal permite-nos identificar a classe de modelos ARIMA, uma extensão dos modelos ARMA apresentados na secção 3.3.2. para processos não estacionários.

De facto no contexto de séries temporais, o operador desfasamento ou *lag operator*, tem um papel importante. Quando aplicado a uma série temporal, o operador de desfasamento, L, atrasa-a ou desfasa-a um período considerado. Por exemplo, quando temos um *lag 1*, $L(z_t) = z_{t-1}$, para $L^2(z_t) = L[L(z_t)] = L(z_{t-1}) = z_{t-2}$. Como este operador é comutativo com a multiplicação, $L(\beta z_t) = \beta L(z_t) = \beta z_{t-1}$, e distributivo relativamente à adição, $L(z_t + x_t) = z_{t-1} + x_{t-1}$, a sua forma geral, com k inteiro é dada por $L^k(z_t) = z_{t-k}$ (Hamilton, 1994).

3.3.2 Modelos estacionários ARMA

Modelos autoregressivos AR(p)

Para os modelos autoregressivos de ordem p, AR (p), o valor Z_t é uma função linear das observações passadas desde Z_{t-1} até Z_{t-p} e de uma função aleatória ε_t, que é o ruído branco. Os modelos AR(1) e AR(p) são dados pelas equações às diferenças estocásticas:

$$\text{AR(1): } Z_t = \emptyset_1 Z_{t-1} + \varepsilon_t \tag{3.24}$$

$$\text{AR}(p): Z_t = \emptyset_1 Z_{t-1} + \emptyset_2 Z_{t-2} + \cdots + \emptyset_p Z_{t-p} + \varepsilon_t \tag{3.25}$$

em que ε_t é um termo independente de Z_{t-k} para $\forall k \geq 1$ e assume-se que os ε_t não estão correlacionados, têm média $E[\varepsilon_t] = 0$ e variância constante $Var[\varepsilon_t] = \sigma^2$. A ordem de autoregressão depende do valor mais antigo p. $\emptyset_i$, $i = 1, \ldots, p$ são os coeficientes de regressão estimados a partir dos dados. O termo (auto)-regressivo significa que existe uma regressão em si mesmo, através de observações anteriores. Por exemplo, se tivermos um modelo AR (1) representado por $Z_t = \emptyset_1 Z_{t-1} + \varepsilon_t$, a observação corrente Z_t é o resultado de uma regressão da observação anterior Z_{t-1} (Brockwell & Richard, 2002; Bisgaard & Kulahci, 2011). Para modelos univariados a série é explicada pelos seus valores passados, nos multivariados ou causais a série é explicada pelos seus valores passados mas também pelos valores passados de outras variáveis.

Modelos de média móvel MA(q)

Nos modelos de média móvel de ordem q, MA(q), o valor presente Z_t é uma função linear de ε_t mais os valores passados dos erros desde ε_{t-1} até ε_{t-q} . Os modelos MA são "médias" dos termos passados e presente do ruído branco (Bisgaard & Kulahci, 2011, p. 60). Este processo é representado pelas equações:

$$\text{MA(1)}: Z_t = \varepsilon_t - \theta \varepsilon_{t-1}, \tag{3.26}$$

$$\text{MA(q)}: Z_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{3.27}$$

Os coeficientes θ_i , $i = 1, \ldots, q$ são parâmetros determinados pelos dados e o sinal negativo é convencional.

Os modelos AR e MA são casos particulares dos modelos mistos ARMA (*p, q*), descritos seguidamente.

Modelos mistos ARMA(p, q)

Finalmente a combinação dos modelos AR(p) e MA(q), são os modelos mistos ARMA(p,q), o valor presente de Z_t é uma função linear dos valores passados da série e dos valores passados dos erros. $\emptyset_i$ e θ_i são os coeficientes de regressão constantes e reais.

$$Z_t = \emptyset_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \emptyset_p Z_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \quad (3.28)$$

3.3.3 As funções de covarância e correlação

A maioria dos processos físicos apresenta uma certa inércia, sem alterações bruscas. Este comportamento dos processos combinado com a frequência da amostragem, em muitas situações, torna as observações consecutivas, correlacionadas entre si. Tal correlação entre as observações consecutivas é chamada de autocorrelação (Bisgaard & Kulahci, 2011) .

Na secção 3.3.2, uma série temporal z_t foi definida como um processo francamente estacionário (ou estacionário de 2ª ordem) se tem uma média e uma variância constantes ao longo do tempo *t*. Quando temos duas variáveis aleatórias X e Y, em geral, define-se a covariância como,

$$Cov(X,Y) = E[\big(X - E(X)\big)\big(Y - E(Y)\big)] \, . \quad (3.29)$$

E a correlação entre X e Y é dada por,

$$Corr\ (X,Y) = \frac{Cov(X,Y)}{\sqrt{\sigma_X^2}\sqrt{\sigma_Y^2}} \quad (3.30)$$

A correlação entre X e Y também pode ser observada através de um gráfico de dispersão.

No contexto de séries temporais, pretende-se definir é a co-variância entre Z_t e Z_{t+k} sendo $E[Z_t] = E[Z_{t+k}] = \mu \ \forall t, k$.

Por conseguinte, tem-se

$$Cov(Z_{t+k}, Z_t) = E[(Z_{t+k} - \mu)(Z_t - \mu)]. \quad (3.31)$$

A função de *autocovariância* pode então ser escrita como uma função do *lag k*

$$\gamma(k) = E[(Z_{t+k} - \mu)(Z_t - \mu)] \quad (3.32)$$

A variância para séries temporais pode ser obtida como $\gamma(0)$. De forma análoga à equação (3.35), a função de autocorrelação (FAC) para séries temporais é dada pela equação

$$\rho(k) = \frac{\gamma(k)}{\sqrt{\gamma(0)}\sqrt{\gamma(0)}} = \frac{\gamma(k)}{\gamma(0)} \quad (3.33)$$

Os coeficientes de autocorrelação assumem valores entre -1 e 1. A função $\rho(k)$ dá-nos as autocorrelações para os diversos *lags*, dando uma visão útil, principalmente se representada graficamente.

Considere-se uma série temporal estacionária modelada através de um processo ARMA. A identificação do modelo específico é determinada examinando a FAC e a função de autocorrelação parcial (FACP). A FAC providencia informação considerável sobre a ordem de dependência quando a série é um processo de média móvel MA. Se o processo é um ARMA ou AR, a FAC por si só diz-nos muito pouco sobre a ordem de dependência. A função de autocorrelação parcial (FACP) entre as variáveis Z_t e Z_{t+k} é mais apropriada, principalmente, para a identificação de um processo AR(p).

A função de autocorrelação parcial mede a correlação (ou autocorrelação) adicional entre Z e Z_{t+k} depois de removido o efeito linear das observações intermédias.

3.4 Séries temporais para dados de contagem

Os modelos ARMA (e ARIMA), anteriormente apresentados, assumem que os dados observados são provenientes de uma distribuição Gaussiana.

Em alguns estudos para dados de contagem, por vezes é ignorado que os dados seguem uma distribuição de Poisson e estes são tratados como gaussianos, utilizando técnicas ARIMA (Saez *et al.*, 1995; Ortíz *et al.*, 1997; Mirón *et al.*, 2010). Estas abordagens falham na adequação do modelo à dinâmica dos dados ou na distribuição natural dos dados de contagem (Brandt & Williams, 2001). Para estudar séries temporais de dados de contagem, Brandt & Williams (2001) propõem o modelo linear autoregressivo de Poisson (PAR – *Poisson*

Autoregressive Model) para dados estacionários. Para dados não estacionários, Brandt & Williams (2000) propõem o modelo PEWMA – *Poisson exponencial weighted moving average model*. Estes dois modelos permitem a sobredispersão dos dados.

Uma abordagem alternativa para lidar com a sobredispersão, o efeito de contágio ou correlação entre os dados de contagem é estimar um modelo generalizado binomial negativo ou modelo GEC – *generalized event count model* (King, 1989).

3.4.1 Modelo de Poisson Autoregressivo – PAR(p)

O modelo autoregressivo permite um efeito causal dinâmico, impondo restrições de estacionariedade. O modelo PAR assume uma distribuição de Poisson para a variável dependente, no entanto, o modelo inclui efeitos aleatórios com distribuição gama para permitir a sobredispersão, sendo a distribuição final dos dados a binomial negativa. Este processo é fundamentado na estatística bayesiana. Trata-se de um modelo de espaço de estados, que pode ser descrito pelas equações seguintes:

Equação de medida: $y_t|m_t \sim P(m_t)$ (3.34)

Equação de transição: $m_t = \sum_{i=1}^{p} \rho_i Y_{t-i} + \left(1 - \sum_{i=1}^{p} \rho_i\right)\mu$ (3.35)

onde $\mu = \exp(X_t\delta)$ indentifica a variável de estado.

Priori conjugada: $m_t|\, Y_{t-1} \sim \Gamma\,(\sigma_{t-1}m_{t-1}, \sigma_{t-1}),\ \mathrm{m_{t-1}} > 0, \sigma_{t-1} > 0$ (3.36)

onde $m_{t-1} = E[y_t|y_{t-1}]$ e $\sigma_{t-1} = Var[y_t|y_{t-1}]$.

Nas equações (3.34) a (3.35) y_t é a variável resposta no tempo *t,* m_t é a média condicionada ao passado do processo linear AR da equação (3.35) sendo o parâmetro Poisson no tempo *t*, p identifica a ordem da componente, ρ_i é o parâmetro autoregressivo para o *lag i*, X_t é o vetor das covariáveis indexadas ao tempo t, $Y_{t-1} = (y_0, y_1, \dots, y_{t-1};\ X_0, X_1, \dots, X_{t-1})$ identifica todo o "passado", ou seja todos os valores observados das variáveis dependentes e independentes. δ é o vetor dos parâmetros de regressão e σ_t é o parâmetro de dispersão. A variável de contagem Y_t é, deste modo, uma função linear dos valores desfasados da variável resposta e uma função exponencial das covariáveis. Além disso, o exponencial da função linear é ponderado para ter um efeito tanto menor quanto maiores forem os parâmetros

autoregressivos. Estas características forçam a que o modelo seja estacionário e impedem a previsão de valores a partir da tendência (Brandt & Williams, 2001; Forgaty & Monogan, 2012).

Quando a série é estacionária, tal pode por exemplo, ser verificado através do correlograma (ou FAC), pode-se adotar um modelo PAR. Se a série for não estacionária sugere-se a aplicação de um modelo PEWMA (Brandt & Williams, 2001), apresentado na secção 3.4.2..

Faz-se notar que a interpretação do modelo PAR(p) difere da dos modelos de Poisson ou Binomial Negativo, uma vez que o primeiro modelo permite identificar o "efeito a curto prazo" de mudanças de X_t sobre o número médio de contagens no tempo t. Por outras palavras, este efeito que é dado pela expressão

$$\left(1 - \sum_{i=1}^{p} \rho_i\right) \exp(X_t \delta)\delta \,, \tag{3.37}$$

identifica o impacto imediato no tempo *t* de uma unidade de mudança na covariável.

Por outro lado, o "efeito a longo prazo" dado por

$$\exp(X_t \delta)\delta \,, \tag{3.38}$$

identifica o efeito total, não considerando as observações passadas. Esta situação ocorre também no modelo de regressão de Poisson *standard*, para o qual "o efeito a curto prazo" e "o efeito a longo prazo" coincidem.

3.4.2 Modelo de Poisson de médias móveis exponencialmente ponderadas - PEWMA

O modelo PEWMA - *Poisson Exponentially Weighted Moving Average*, fundamentado nos estudos de Harvey & Fernandes (1989), trata-se de um modelo de espaço de estados que assume um processo latente dinâmico subjacente que é contínuo e este processo contínuo é mapeado sobre a variável observada discreta. O valor médio varia ao longo do tempo e na interpretação do modelo dinâmico, o parâmetro de interesse descreve a quantidade de variação ao longo do tempo da média (Brandt & Williams , 1998;2000).

Os proponentes deste modelo consideram que a regressão de Poisson clássica é um caso especial do modelo PEWMA e o impacto das covariáveis tem a mesma interpretação dos coeficientes do modelo estático de regressão de Poisson ou binomial negativo.

Os valores observados no tempo *t* seguem uma distribuição marginal de Poisson,

$$\Pr(y_t|\mu_t) = \frac{\mu_t^{y_t} e^{-\mu_t}}{y_t!} \tag{3.39}$$

onde o parâmetro μ_t identifica a taxa média não observada para a contagem de tempo t.

Este processo tal como o seu homólogo PAR é fundamentado na estatística bayesiana. O modelo pode ser descrito pelas equações seguintes :

Equação de medida: $y_t|\mu_t \sim P(\mu_t)$ e $\mu_t = \mu_{t-1}^{*} \exp(X_t\delta)$ (3.40)

Equação de transição: $\mu_t = e^{r_t}\mu_{t-1}\eta_t$ (3.41)

Priori conjugada: $\mu_{t-1}^{*} \sim \Gamma(a_{t-1}, b_{t-1})$ (3.42)

Nas equações acima δ é o vetor dos coeficientes, X_t é o vetor das variáveis explicativas e μ_{t-1}^{*} é a componente de variação no tempo. Esta componente é estimada pelo filtro de Kalman e representa as contagens observadas à priori no tempo $t-1$, ou seja, é uma média suavizada das observações anteriores. O mecanismo estocástico para a transição nas séries temporais do tempo $t-1$ para o tempo t é dado pela relação entre μ_{t-1} e μ_t. A dinâmica da média é então descrita pela equação (3.41), onde η_t segue a distribuição beta $\beta(\omega a_{t-1}, (1-\omega)a_{t-1})$ na qual o parâmetro ω captura o desconto nas observações para o cálculo da média.

As variáveis η_t e r_t parametrizam a taxa de crescimento no periodo t, assegurando que $\mu_t > 0$. Das propriedades da distribuição beta, conclui-se que E[η_t] = ω para todo *t*.

3.4.3 Modelo *lagged* Poisson

O modelo de regressão de Poisson *standard*, tal como já referido, pode ser representado pela parametrização

$$\mu = \exp(X_t\delta).$$

No entanto, este modelo assume erradamente que as observações são independentes. Em vários estudos bioclimáticos, se os dados são uma série temporal, é incluída uma variável dependente desfasada (Pollins, 1996) ou uma variável preditora desfasada (Almeida *et a*l, 2010; Peng & Dominici, 2011; Monteiro *et al.*, 2012).

Como alternativa aos modelos PAR(p) e PEWMA, surge então o modelo de regressão *lagged* de Poisson (ou Poisson com desfasamento nos regressores), que se pode apresentar da seguinte forma

$$Y_t \sim P\ (\mu_t) \text{ onde } \mu_t = \exp\ (\ X_t\delta + \ \rho z_{t-1}) \tag{3.43}$$

O modelo de *lagged* Poisson assume que o efeito de um aumento em uma unidade na variável $x_{t-\ell}$ num determinado dia, **estende-se até *k* dias para o futuro**. Sendo este modelo mais adequado à realidade, assumindo o descrito na frase anterior, é necessário termos disponíveis os dados diários da variável $x_{t-\ell}$ (Peng & Dominici, 2011).

Em termos mais genéricos, para um modelo que envolva observações da covariável até os *p* dias anteriores, a média pode ser escrita como

$$\exp\ (X_t\delta + \rho y_{t-1} + \cdots + \ \rho y_{t-p}\ ) \tag{3.44}$$

De acordo com os autores Brandt & Williams (2001), a desvantagem deste modelo passa por implicar uma taxa de crescimento não-nula para a média condicional. O coeficiente ρ deixa de representar um coeficiente de correlação e passa a representar um valor para a taxa de crescimento. Este tipo de modelos são então mais adequados para dados não estacionários. Os modelos PEWMA e *lagged* Poisson assumem que as observações são dependentes.

CAPÍTULO IV

4 DESCRIÇÃO DOS DADOS E RESULTADOS

4.1 Área de estudo

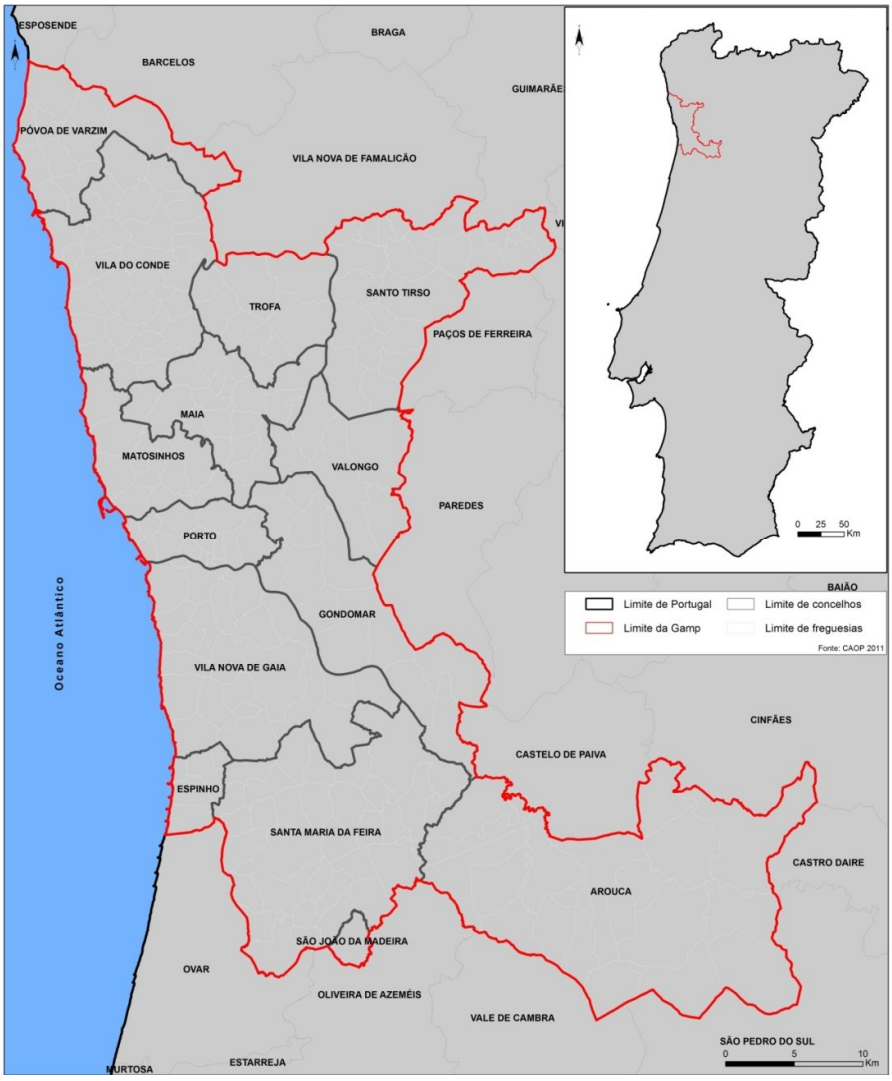

Figura 3 Limites administrativos da Grande Área Metropolitana do Porto

A GAMP situa-se no Litoral Norte de Portugal e agrupa 16 concelhos numa área de 2089 km². Tem uma densidade populacional próxima de 1098 hab/km², que totalizava 2294741 habitantes, em 2011. A GAMP compreende a área do Grande Porto e de Entre Douro e Vouga. A área do Grande Porto é constituída pelos concelhos de Espinho, Gondomar, Maia, Matosinhos, Porto, Póvoa de Varzim, Santo Tirso, Trofa, Valongo, Vila do Conde e Vila Nova de Gaia. A área de Entre Douro e Vouga compreende os concelhos de Arouca, Oliveira de Azeméis, Santa Maria da Feira, São João da Madeira e Vale de Cambra. No ano de 2007 a GAMP era constituída apenas por catorze concelhos, Figura 3, ainda não agrupava os concelhos Oliveira de Azeméis e Vale de Cambra.

Dados da saúde e ambientais

Os dados estatísticos para a concretização deste projeto foram fornecidos pelo Departamento de Geografia da Faculdade de Letras da Universidade do Porto, local de realização do estágio. De forma indireta estes dados foram cedidos pelas seguintes entidades:

- a base de dados sobre a mortalidade registada entre os anos 2002 e 2007, foi cedida pelo Instituto Nacional de Estatística e inclui registos diários de mortalidade e de internamentos (2002-2007), realizados em quatro hospitais da Grande Área Metropolitana do Porto (GAMP);

- os dados diários meteorológicos, relativos à temperatura e humidade relativa, foram cedidos pelo Observatório Meteorológico da Serra do Pilar;

- e os dados da concentração de poluentes atmosféricos, ozono e PM_{10} foram cedidos pela Agência Portuguesa do Ambiente.

No nosso estudo, uma outra variável importante é a temperatura aparente (HI – *Heat index)* que representa o efeito da exposição à temperatura que é sentida, normalmente, nas estações mais quentes. O índice de temperatura aparente[6] foi calculado de acordo com a

[6] Existem outras fórmulas para calcular a temperatura aparente. Por exemplo, Kalkstein & Valimont (1986, p.843) , apresentaram a seguinte equação: $AT = -2.653 + (0.994\ T_a) + 0.368(T_d)^2$, onde AT=temperatura aparente (ºC), T_a= temperatura do ar (ºC) e T_d= orvalho (ºC).

fórmula apresentada por Rothfusz (1990), onde T é a temperatura do ar em ºC e R é a humidade relativa em %:

$$HI = -42.379 + (2.04901523 * T) + (10.14333127 * R) - (0.22475541 * T * R) - (6.83783 * 10^{-3} * T^2) - (5.4481717 * 10^{-2} * R^2) + (1.22874 * 10^{-3} * T^2 * R) + (8.5282 * 10^{-4} * T * R^2) - (1.99 * 10^{-6} * T^2 * R^2)$$

Os métodos estatísticos utilizados foram variados e incluem os modelos lineares generalizados, os modelos aditivos generalizados e os modelos de regressão de séries temporais. Nos modelos apresentados neste projeto, a variável resposta são dados de contagem, uma vez que identifica o número de óbitos. As variáveis explicativas são a temperatura aparente, a concentração de ozono e a concentração de PM_{10}.

A variável de interesse, a mortalidade, é presumivelmente afetada pela temperatura aparente e por agentes de poluição.

Os modelos foram ajustados pela sazonalidade, com a análise limitada às estações de verão que compreende os meses de junho, julho e agosto. A análise foi efetuada com recurso ao software *R* (*The Comprehensive R Archive Network*: http://cran.r-project.org/).

4.2 Análise descritiva

Numa fase inicial do trabalho, fez-se uma análise descritiva das quatro principais variáveis envolvidas no estudo, para os verões dos anos de 2002 a 2007. A Figura 4 e a Tabela 2 apresentam as principais estatísticas descritivas para os verões dos 6 anos envolvidos no estudo.

Os gráficos *boxplots* apontam observações aberrantes (*outliers*) em todas as variáveis. Comparando os verões dos 6 anos em estudo, o maior número de óbitos foi registado em 2006, com um máximo de 59 e um mínimo de 9 óbitos (M = 26.75, DP = 8.38), no entanto, o ano de 2003 registou valores médios mais elevados (M =28.41, DP = 6.37), com um máximo de 49 e um mínimo de 15 óbitos. O ano de 2003 foi um dos últimos anos afetados pelas ondas de calor.

A temperatura aparente apresentou o valor médio mais alto (M = 29.20, DP = 6.48) no ano de 2006, o valor máximo de 55.10 ºC no ano de 2005 e o valor mínimo mais elevado de 20.70 º C no ano de 2004.

A concentração das partículas PM_{10} registou no ano de 2005 o valor médio mais elevado (M = 51.33, DP = 24.36), assim como o valor mínimo mais elevado, 15 $\mu g/m^3$, enquanto o seu valor máximo, 126 $\mu g/m^3$, foi registado no ano de 2002. A concentração do poluente ozono (O_3) apresentou valores médios mais elevados no ano de 2006 (M = 54.22; DP = 20.85), com um valor máximo de 92 $\mu g/m^3$ no ano de 2003 e um valor mínimo mais elevado, 23 $\mu g/m^3$, no ano de 2005.

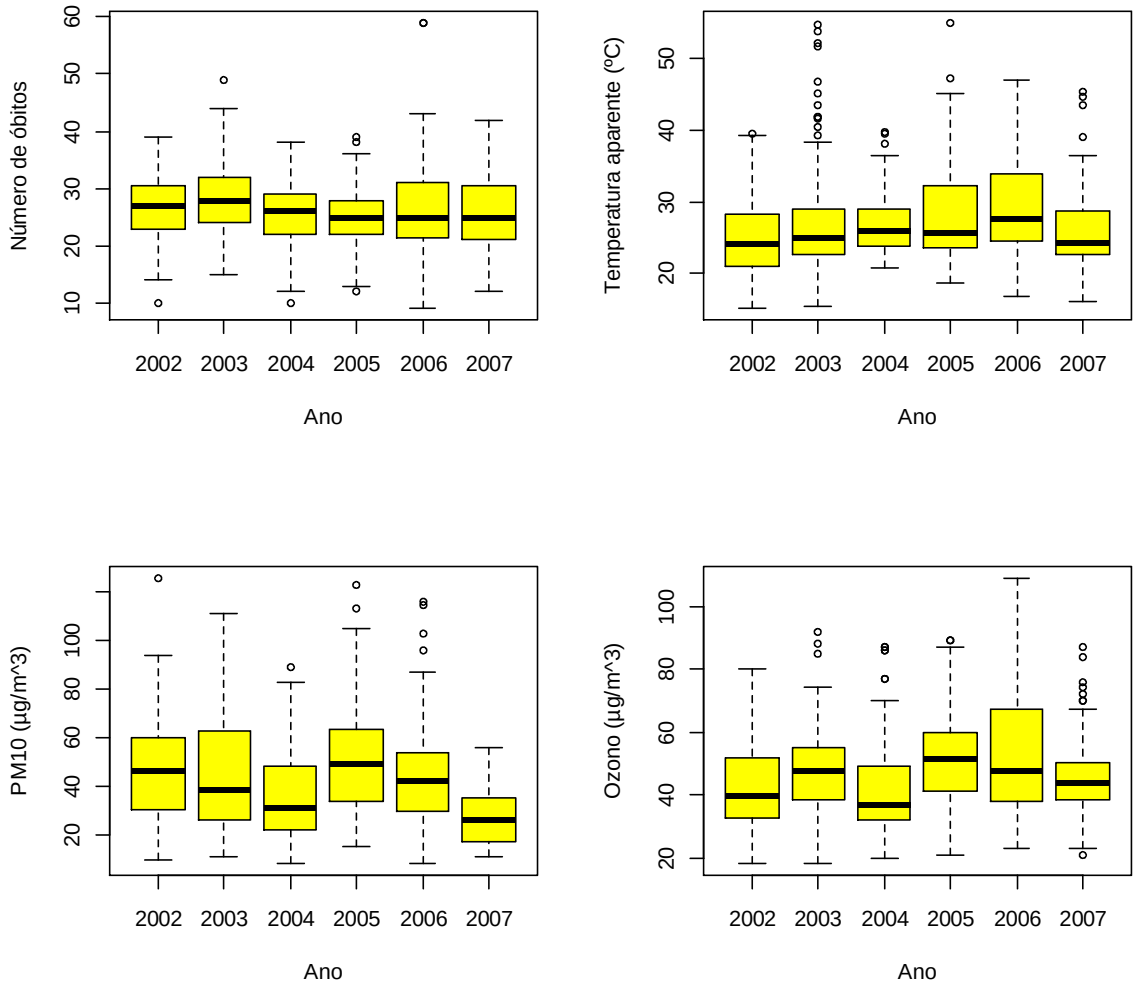

Figura 4 Gráficos *boxplot* para a mortalidade, a temperatura aparente, as PM_{10} e o ozono, nos verões de 2002 a 2007

Tabela 2 Estatística descritiva para as variáveis mortalidade, temperatura aparente, concentração de PM_{10} e concentração de ozono

		2002	2003	2004	2005	2006	2007
Numero de óbitos	Média	26.61	28.41	25.79	24.66	26.75	25.59
	desvio-padrão	5.35	6.37	5.46	5.31	8.38	6.83
	Mínimo	10	15	10	12	9	12
	Máximo	39	49	38	39	59	42
Temperatura aparente (∘C)	Média	24.77	27.73	26.87	28.57	29.20	26.16
	desvio-padrão	5.13	8.52	4.23	6.93	6.48	5.53
	Mínimo	15.10	15.40	20.70	18.80	16.80	16.20
	Máximo	39.60	54.80	39.80	55.10	47.10	45.50
$PM_{10}(\mu g/m^3)$	Média	46.66	45.36	51.33	44.85	27.11	27.11
	desvio-padrão	20.20	24.75	24.36	21.37	10.84	10.84
	Mínimo	10	11	15	8	11	11
	Máximo	126	111	123	116	56	56
$O_3(\mu g/m^3)$	Média	42.66	47.99	41.82	54.22	45.75	45.75
	desvio-padrão	14.29	13.34	15.68	20.85	13.28	13.28
	Mínimo	18	18	20	23	21	21
	Máximo	80	92	87	59	87	87

Para as análises que se seguem foi utilizada uma amostra restrita ao verão de 2006, uma vez que este foi o ano que apresentou o maior número de óbitos, de entre os verões dos anos 2002 a 2007. Nesta análise exploratória analisamos a relação entre as variáveis através do gráfico de dispersão e do teste de correlação de *Pearson*.

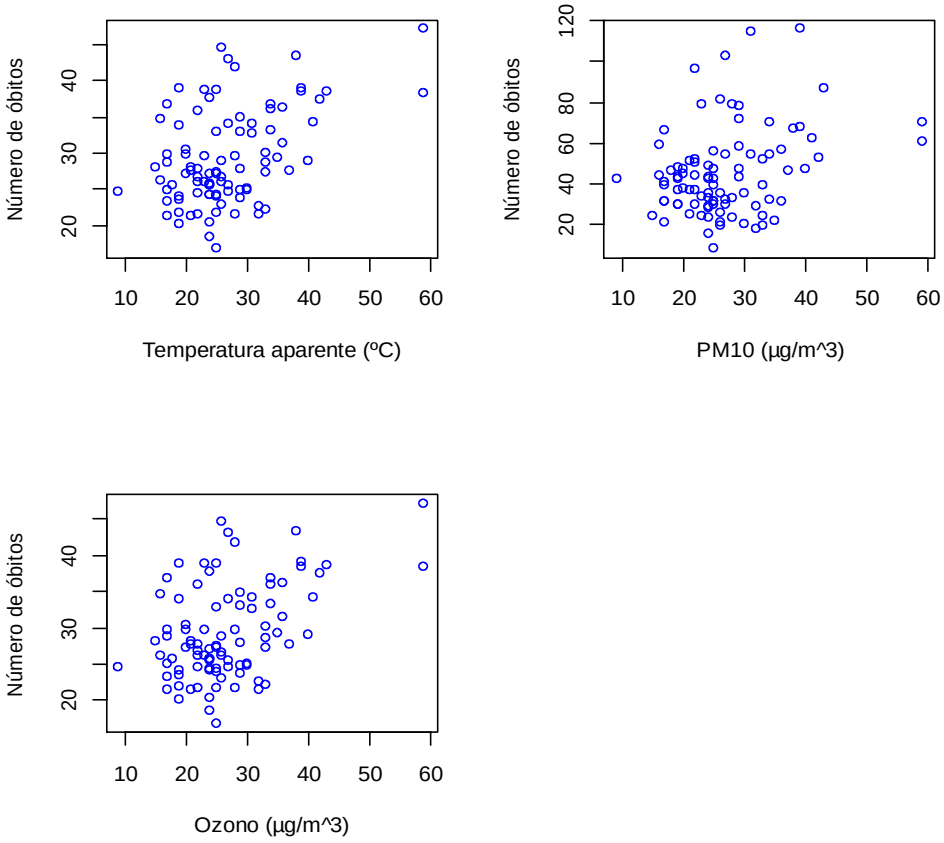

Figura 5 Diagramas de dispersão da relação da mortalidade com as variáveis temperatura aparente, concentração de ozono e concentração de PM_{10}.

Tabela 3 Coeficiente de correlação de *Pearson*

	Número de óbitos	**Concentração de ozono**	**Concentração de $PM10$**
Concentração de ozono	.44** (<.001)		
Concentração de $PM10$	.30 *(<.01)	.55** (<.001)	
Temperatura aparente	.45 **(<.001)	.42** (<.001)	.62**(<.001)

Os gráficos da Figura 5 sugerem uma associação positiva entre o número de óbitos e as outras variáveis. A Tabela 3 indica que a mortalidade está positivamente correlacionada com a concentração de PM_{10} ($r = .30, p < .01$), com a concentração de ozono ($r = .44, p<.001$) e com a temperatura aparente ($r = .45, p < .001$).

Na Figura 6, estão representados os gráficos das séries temporais das variáveis em estudo. Esta figura sugere que a mortalidade, a temperatura aparente (ºC), a concentração do ozono ($\mu g/m^3$) e a concentração das partículas PM_{10} ($\mu g/m^3$) comportam-se ao longo do tempo de forma idêntica. Os gráficos desta figura indicam que, em todas as variáveis há um grande aumento, um "pico", no mês de julho que parece estar de acordo com a onda de calor registada entre os dia 7 e 18 deste mês. Notando-se outro aumento, mas menos intenso, durante o mês de agosto, que registou uma onda de calor de 2 a 11 de agosto.

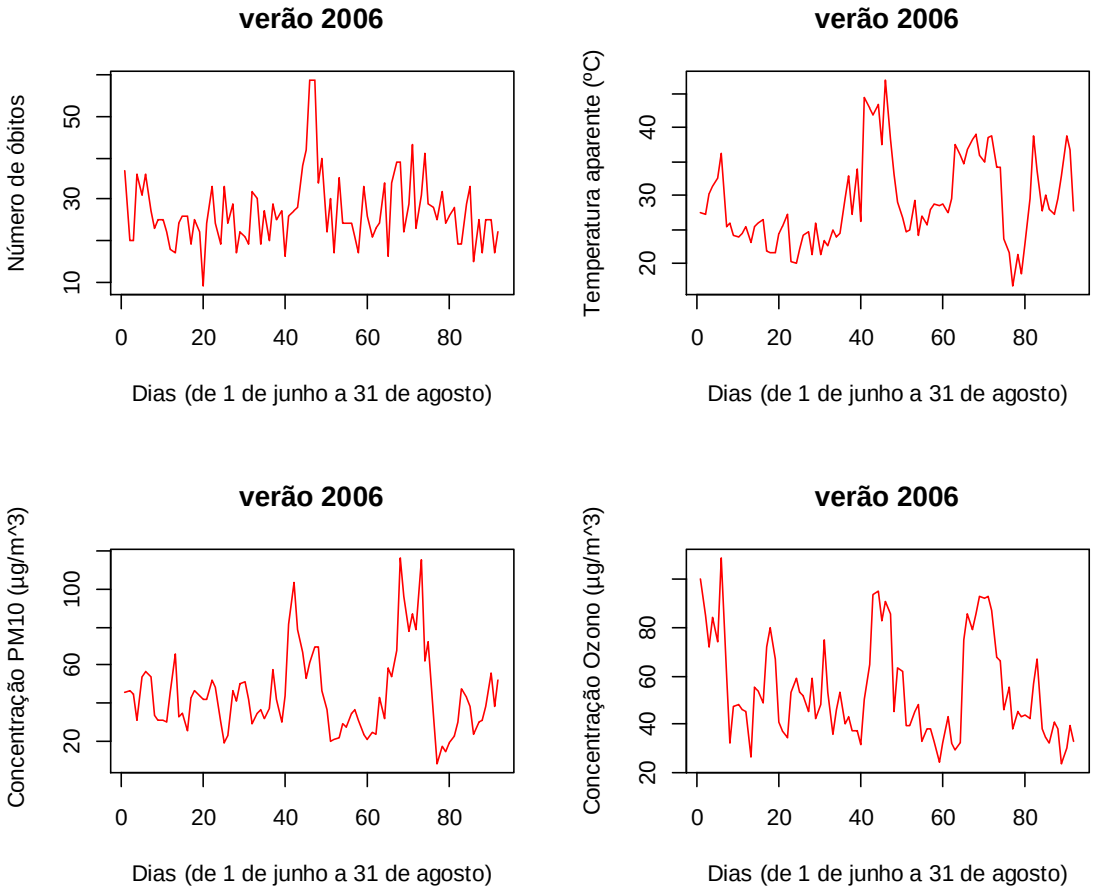

Figura 6 Séries temporais das variáveis mortalidade, a temperatura aparente, concentração de ozono e concentração de PM_{10}.

4.3 Resultados preliminares da modelação

4.3.1 Modelos aditivos generalizados

Por se tratar de um modelo clássico mais abrangente, nesta secção iremos dar maior ênfase ao MAG que surge como uma alternativa para a modelação de relações não-lineares que apresentem uma forma não definida, baseando-se em funções não-paramétricas, designadas por curvas de suavização (Hastie & Tibshirani, 1990). De forma similar aos investigadores Basu, *et al.* (2008), Almeida *et al*. (2010), Peng & Dominici (2011) e muitos outros estudos na área da Bioestatística, que estudam os efeitos de variáveis ambientais, como a temperatura aparente, temperatura mínima e máxima e os efeitos da poluição atmosférica, na saúde dos habitantes de grandes centros urbanos, utilizaram os modelos aditivos generalizados, com a função de ligação *quasi*-Poisson.

Nesta análise, utilizamos um modelo aditivo generalizado de regressão *quasi*-Poisson tendo como variável dependente o número de óbitos e como variáveis independentes a temperatura aparente, o dia da semana (variável qualitativa), a concentração de PM_{10} e a concentração de ozono. O dia da semana foi incluído para controlar a tendência semanal.

Há evidências, tanto teóricas como empíricas, de que variáveis da saúde, como a mortalidade e a morbilidade, são influenciadas por fatores sazonais e outros fatores de tendencia temporal. Neste caso, é aconselhável fazer o ajuste das tendências temporais. Por exemplo, o número de atendimentos hospitalares é mais alto durante a semana do que ao fim de semana. Para ajustar esta tendência, pode-se acrescentar uma variável qualitativa para o dia da semana (Tadano, 2007). Também a concentração de poluentes atmosféricos e os fatores metereológicos variam ao longo do ano, assim sendo, a sazonalidade também é um fator importante a considerar. Para ajustar esta tendência temporal é frequente utilizar um *spline,* uma vez que fornece uma apoximação do comportamento das funções com alterações locais e bruscas (Chapra & Canale , 1987; Peng & Dominici, 2011).

Os MAG´s ajustados são modelos semi-paramétricos, com uma função *spline* aplicada à temperatura aparente, à concentração de PM_{10} e à concentração de ozono. Para controlar a tendência temporal e a sazonalidade e deste modo avaliar a robustez dos resultados obtidos,

foi incluído o suavizador *spline* cúbico com graus de liberdade aproximadamente igual ao número de meses do período em estudo na covariável temperatura aparente (Kelsall, JM, Zeger, & Xu, 1997).

Na Figura 7 estão desenhados os *splines* cúbicos para o modelo de regressão que estuda a relação entre o número de óbitos e a temperatura aparente para 1, 3, 4 e 8 graus de liberdade, onde se pode observar um aumento diretamente proporcional da flexibilidade da curva de suavização com o aumento dos graus de liberdade.

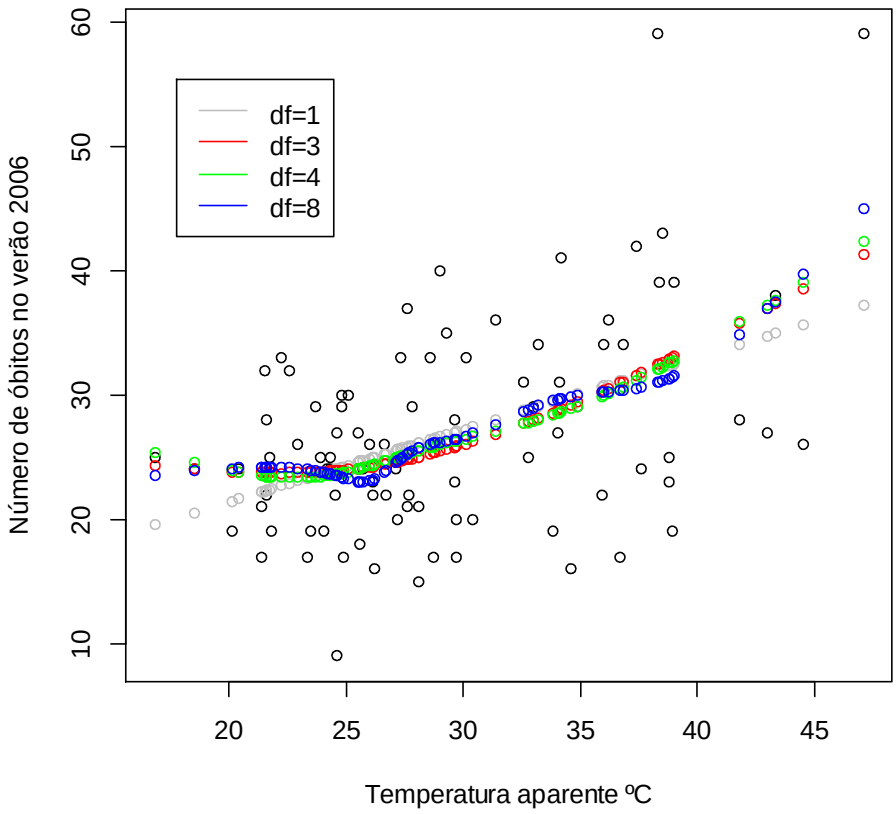

Figura 7 *Splines* cúbicos para o modelo de regressão *mortalidade total ~ temperatura aparente*, ajustado para 1, 3, 4 e 8 graus de liberdade.

O modelo com 4 graus de liberdade sugeriu melhores resultados. Quando utilizamos 8 graus de liberdade os valores dos coeficientes estimados são sobrestimados, tal como refere (Almeida, *et.al.* 2010).

Foram analisados dois modelos que integram a temperatura aparente como covariável: a temperatura aparente no dia atual (o modelo com *lag 0)* e a temperatura aparente no dia anterior (modelo com *lag 1*). O modelo com *lag 1* sugeriu melhores resultados.

A Tabela 4 mostra os resultados dos modelos em estudo. Para a GAMP no verão de 2006, pelo aumento de 1ºC da temperatura aparente a mortalidade aumenta em 2.2 % (95% IC: 1.4-3.0). Pelo aumento de 1ºC da temperatura aparente, observa-se um aumento na mortalidade de 1.7 % (95% IC: 0.7-2.7), 1.7% (95% IC: 0.5-2.5) e 1.5 % (95% IC: 0.8-2.6), quando controlado o efeito da concentração das PM_{10}, da concentração do ozono, ou da concentração dos dois poluentes em simultâneo, respetivamente.

Tabela 4 % de aumento (Intervalo de Confiança a 95%) na mortalidade pelo aumento em 1º C em média da temperatura aparente (*lag* 1), ajustado pelos poluentes (*lag* 0) individuais e em conjunto, na GAMP[a].

Mortalidade	***% aumento***	***IC a 95%***
Modelo base: s(temperatura aparente + *df* = 4)+ dia da semana +	2.2	1.4-3.0
+ s(PM10)	1.7	0.7-2.7
+s(ozono)	1.7	0.5-2.5
+s(PM10)+s(ozono)	1.5	0.8-2.6

[a] Modelo ajustado pelo dia da semana

O estudo realizado por Almeida *et al.* (2010) que investigaram o efeito da temperatura aparente na mortalidade, mostrou que para a cidade do Porto, durante os meses de abril a setembro de 2000-2004, pelo aumento em 1ºC da temperatura aparente a mortalidade para todas as idades e por todas as causas aumentou 1.5 % (95% IC: 1.0 - 1.9). O aumento no valor esperado da mortalidade no nosso estudo em comparação com o valor encontrado por estes investigadores é justificado pela nossa concentração no verão de 2006 que registou 3 ondas de calor, neste período.

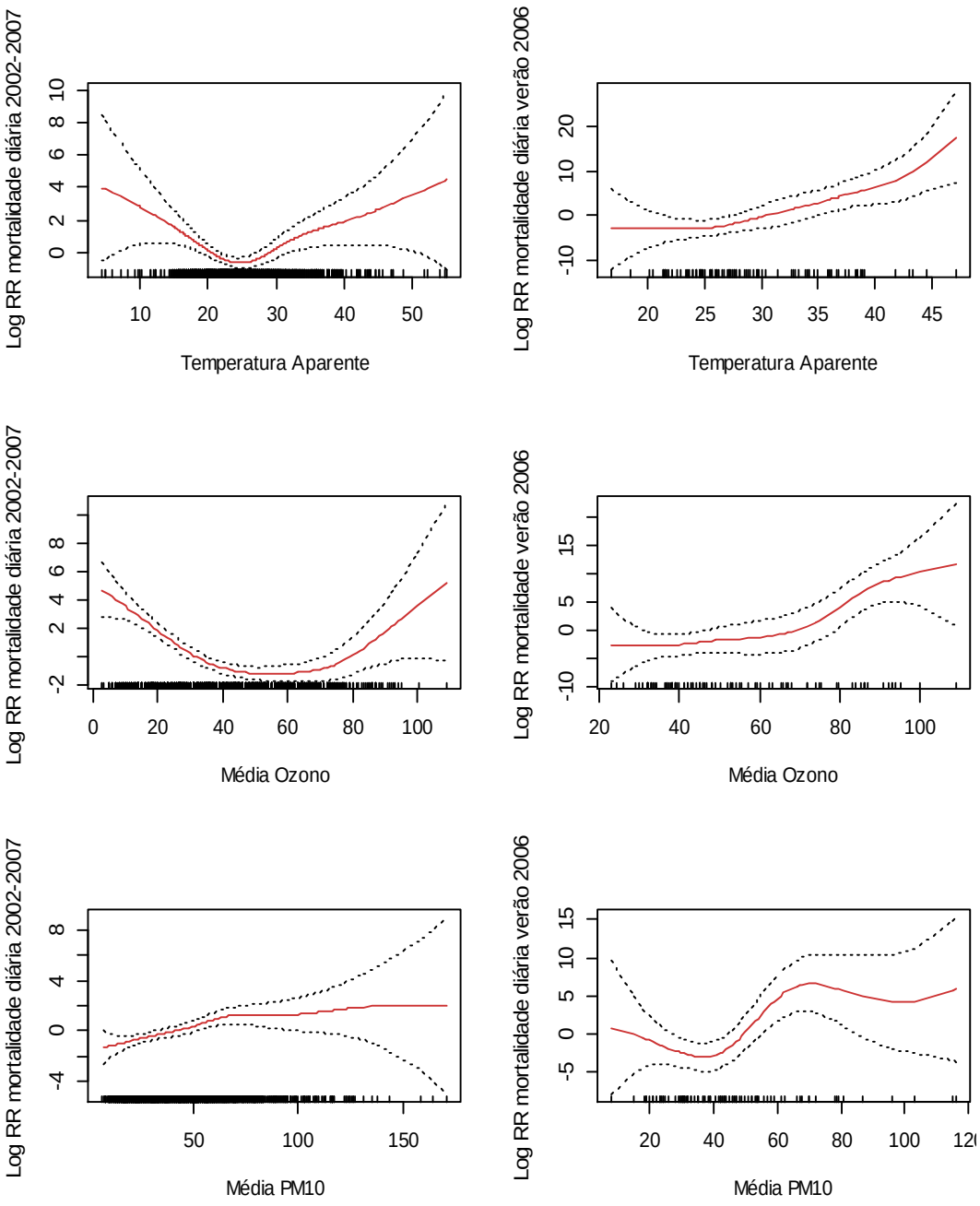

Figura 8 Curvas de suavização para os modelos em estudo

Na Figura 8 estão representadas as funções de suavização para os modelos em estudo. O gráfico da mortalidade *versus* temperatura aparente apresenta em geral uma forma em V (imagem do lado esquerdo referente aos registos diários do ano de 2006), notando-se um aumento da mortalidade a partir dos 25 ºC. Para os dados do verão de 2006, a imagem do lado direito sugere um efeito quase linear da temperatura aparente na mortalidade. É notável para

todos os modelos, a diferença entre a representação gráfica e suavizada dos dados, para os registos diários ao longo do ano de 2006, quando comparada com a representação dos dados apenas para o verão de 2006.

4.3.2 Modelos para séries temporais

Para as análises que se seguem também foram utilizados os dados do verão de 2006. O correlograma da FAC da Figura 9 sugere que a série não é aleatória, os valores desfasados estão correlacionados. Os limites de confiança aproximados de 95% são dados por $\pm 1.96 / \sqrt{n}$, no nosso caso os limites de confiança são aproximadamente $\pm 1.96 / \sqrt{92} \approx 0.20$ e podemos notar que 3 de entre as 92 observações estão fora destes limites, o que nos leva a concluir que não há evidência para rejeitar a hipótese de que as observações não são independentes.

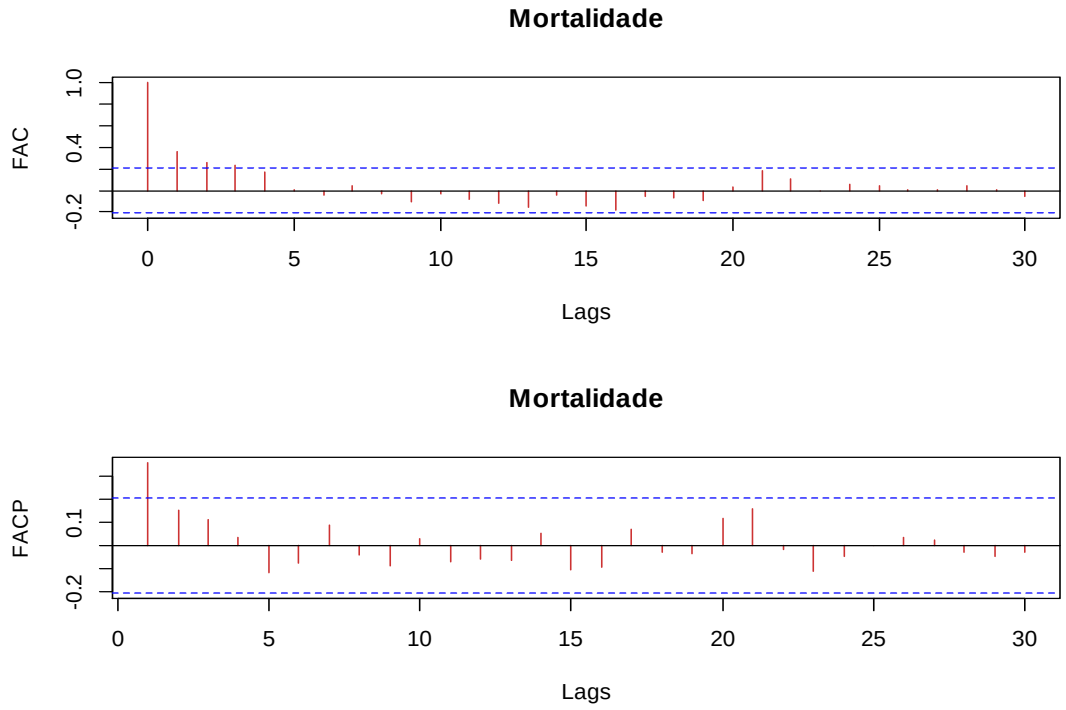

Figura 9 FAC e FACP para a mortalidade

Pela análise da Figura 9, pode-se concluir que a FAC decresce exponencialmente para zero. Relativamente à FACP, esta anula-se a partir de um intervalo maior que 1 (de facto, a

partir do *lag 1*, as correlações parciais caem no intervalo definido como não significativamente diferente de zero). A combinação destes dois gráficos permitir-nos-iam sugerir um modelo autoregressivo de primeira ordem AR(1) para os dados da mortalidade. No entanto, como se tratam de dados de contagem, tal como se explica na Secção 3.4, iremos recorrer aos modelos PAR(p) ou PEWMA. Mais especificamente, como o gráfico para a mortalidade da Figura 6 mostra que esta série é estacionária para a média e não estacionária para a variância, iremos recorrer ao modelo PEWMA.

4.4 Comparação dos modelos

Para analisarmos os efeitos específicos da distribuição e da dinâmica da mortalidade, iremos restringir o nosso estudo de comparação a quatro modelos plausíveis:

1. Modelo linear generalizado de Poisson
2. Modelo linear generalizado Binomial Negativo
3. PEWMA
4. Modelo *Lagged* Poisson.

Para simplificação da análise avaliamos a relação da mortalidade apenas com uma variável independente, a temperatura aparente. No que se refere à interpretação do valor esperado dos modelos, Brandt & Williams (2000:2001) referem que para o modelo de regressão de Poisson, o modelo de regressão binomial negativa e o modelo PEWMA, a mudança de percentagem no valor esperado pela alteração em uma unidade no regressor é dada por

$$100.\,(exp(\beta) - 1) \tag{4.1}$$

Os modelos apresentados no Capítulo 3 permitem estimar o valor esperado do aumento da percentagem da mortalidade, pelo aumento de um 1ºC da temperatura aparente, através das equações da média condicionada (Brandt & Williams, 2001), nomeadamente tem-se

$$\text{GLM:}\quad E[y_t|X_t] = \log(\mu_t) = \exp(\beta_0 + X_t\beta_1) \tag{4.2}$$

onde X é a temperatura aparente.

$$\textit{Lagged}\ \text{Poisson:}\quad E[y_t\,|X_t\,, y_{t-1}] = \exp(\beta_0 + X_t\beta_1 + \rho y_{t-1}) \tag{4.3}$$

$$PEWMA: \quad E\left[y_t \left| X_t, y_{t-1}, \ldots, y_{t-p}\right.\right] = \mu_{t-1}^{*} \exp\left(\beta_0 + X_t\beta_1 + r_t\right)\eta_t \qquad (4.4)$$

Os parâmetros estimados (e o respectivo erro padrão) para estes modelos estão apresentados na Tabela 5. Para cada modelo estimamos os parâmetros pelo método da máxima verosimilhança. Todas as análises efetuadas indicam que o efeito do aumento da temperatura aparente na mortalidade é significativo.

Tabela 5 Resultados dos modelos Poisson, Binomial Negativo, PEWMA e *Lagged* Poisson

	Poisson	**Binomial negativo**	**PEWMA**	***Lagged* Poisson**
Constante	2.668	2.680	--------	2.550
Coeficiente estimado	0.021 (0.003)	0.020 (0.004)	0.019 (0.005)	0.025 (0.003)
ω			0.68 (0.07)	
θ – parâmetro de dispersão	1	29.61 (9.02)	--------	1
N	92	92	92	89
AIC	647.37	623.09	630.96	606.16
Log-likelihood	-321.68	-308.54	-314.48	-301.08

O primeiro modelo estimado é o modelo de regressão de Poisson, que sugere que pelo aumento de 1ºC da temperatura aparente a mortalidade aumenta 2.1%. Em estudos similares com MLG´s utilizando a regressão de Poisson, Monteiro, *et.al.* (2012) analisaram o efeito da temperatura aparente na mortalidade total, na cidade do Porto, em Portugal, durante a onda de calor no mês de julho de 2006, concluíndo que pelo aumento de 1º C na temperatura aparente, a mortalidade aumenta 0.6 % quando controlado o efeito do ozono e aumenta 0.4%, quando controlado o efeito das partículas PM_{10}.

O segundo modelo é o modelo de regressão binomial negativa. O resultado para esta análise indica que, pelo aumento de 1ºC da temperatura aparente a mortalidade aumenta 2.0%.

O terceiro modelo é o modelo Poisson exponencial de média móvel ponderada, PEWMA, que sugere que pelo aumento de 1ºC da temperatura aparente a mortalidade aumenta 1.9 %.

O quarto modelo é o modelo de regressão *Lagged* Poisson, que considera a temperatura aparente dos dias anteriores como variável explicativa. A Tabela 6 apresenta os coeficientes estimados da temperatura aparente para os modelos dos primeiros 4 *lags*. Para o modelo do *lag* 0, como seria de esperar, o coeficiente estimado é igual ao coeficiente estimado para o modelo de regressão de Poisson. O coeficiente estimado para o modelo do *lag* 1 tem um valor ligeiramente menor em relação ao modelo de com *lag* 0. No entanto, os valores dos coeficientes dos modelos seguintes com *lag 2* e com *lag 3*, são crescentes. O valor apresentado na Tabela 5 corresponde ao coeficiente estimado para o *lag* 3, no qual se verifica a percentagem mais alta de mortalidade total, 2.5%. De entre os modelos com *lags* singulares na regressão *lagged* Poisson, o modelo com *lag 3*, também foi o que apresentou o menor *AIC*.

Tabela 6 Resultados para os modelos singulares da *Lagged* Poisson

Lags	***Betas estimados***	***Erro padrão***	*AIC*
Lag 0 Temperatura aparente	0.02086	0.00300	647.37
Lag 1 Temperatura aparente	0.01979	0.00302	643.41
Lag 2 Temperatura aparente	0.02165	0.00302	628.69
Lag 3Temperatura aparente	0.02498	0.00303	606.16

Os coeficientes estimados têm valores crescentes aproximadamente até ao *lag* 3, a partir do qual os valores dos coeficientes estimados caem rapidamente até se aproximarem de zero, Figura 10. Por outras palavras, a função do efeito do aumento de 1ºC na temperatura aparente na mortalidade estende-se até 4 dias, contando o primeiro dia como a observação no *lag* 0, de forma positiva, suave e crescente, começando a perder força rapidamente a partir desse dia.

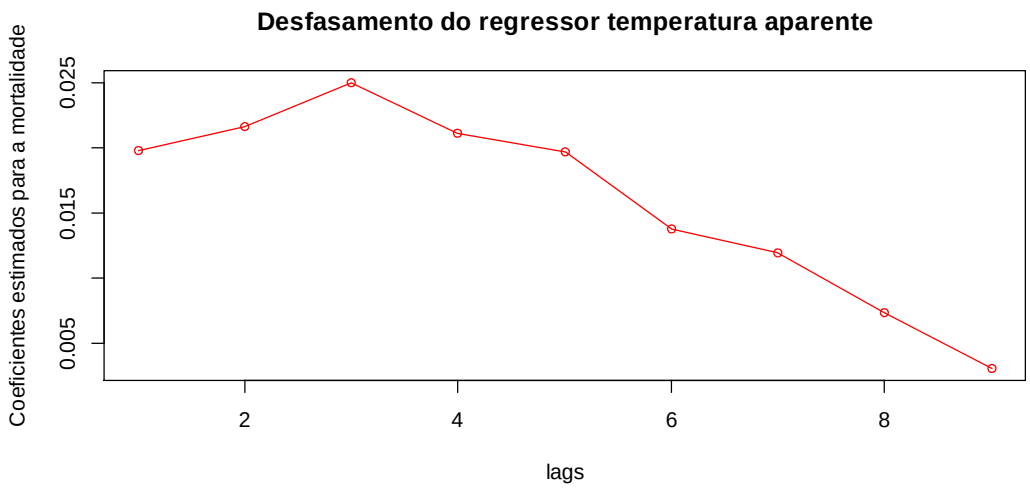

Figura 10 Coeficientes estimados para os primeiros 9 *lags* da temperatura aparente da regressão *Lagged* Poisson

CAPÍTULO V

5 CONCLUSÕES E TRABALHOS FUTUROS

Conclusões

As ondas de calor afetam de forma significativa a saúde humana, provocando principalmente danos circulatórios, respiratórios, neuronais e consequentemente levam a um aumento da taxa de mortalidade.

Referenciando a máxima de George Box, *"Essencialmente, todos os modelos são errados mas alguns são úteis",* no nosso estudo, os diferentes modelos estatísticos sugerem valores esperados para a taxa de mortalidade muito próximos, que variaram de 1.5% até 2.5 %. O que nos leva a concluir que os diferentes modelos estatísticos com maior ou menor eficiência, aproximam-se do valor esperado real, permitindo ao investigador maior flexibilidade e criatividade na análise estatística.

Os modelos que pressupõem a independência das observações, que são o modelo de regressão de Poisson, o modelo aditivo generalizado e o modelo de regressão binomial negativo, sugeriram que pelo aumento de 1ºC da temperatura aparente a mortalidade aumenta 2.1%, 2.2% e 2.0%, respetivamente. Estes valores são valores muito próximos sugerindo que podemos utilizar qualquer um destes modelos com algum grau de confiança, para dados de contagem com as características dos dados utilizados neste estudo. Os modelos que pressupõem que as observações são dependentes, ou estão correlacionadas, que são o modelo PEWMA e o modelo de regressão *lagged* Poisson, sugeriram que pelo aumento de 1ºC da temperatura aparente a mortalidade aumenta 1.9% e 2.5%, respetivamente. Estes valores não são tão aproximados como os valores dos modelos referidos anteriormente, no entanto, a diferença reside na dinâmica do modelo *lagged*.

De acordo com Brandt & Williams (1998; 2000) e considerando que existe dependência temporal, o modelo *lagged* Poisson é sempre mais eficiente que os modelos de regressão de

Poisson e binomial negativo. Os modelos *lagged* para eventos de contagem são geralmente menos eficientes que o PEWMA. Utilizando o critério de AIC, neste estudo, de facto, o modelo *lagged* Poisson mostrou mais eficiência que os modelos de regressão de Poisson e binomial negativo. No entanto, também mostrou melhor eficiência que o PEWMA, contrariando neste ponto as análises efetuadas por Brandt & Williams.

Para os modelos estáticos de regressão de Poisson, de regressão binomial negativa e PEWMA a interpretação do valor esperado é do efeito imediato da temperatura aparente na mortalidade (Brandt & Williams, 2001; Forgaty & Monogan, 2012). A omissão da dinâmica nestes modelos leva a uma subavaliação da magnitude da temperatura aparente, uma vez que não leva em conta a dinâmica de um aumento futuro no valor esperado da mortalidade.

Uma das desvantagem dos modelos de regressão de Poisson, binomial negativo e PEWMA é que não permitem estimar o efeito para o futuro, o valor esperado para a variável resposta é estático. Se um modelo dinâmico tem realmente precisão, então a forma funcional de obter o resultado esperado do modelo não é fundamentalmente correta, se utilizarmos um modelo estático (Forgaty & Monogan, 2012; Brandt & Williams, 2001). Modelos de contagem de eventos estáticos prevêm uma média constante, em vez de capturar os efeitos ao longo do tempo no valor esperado da variável dependente.

A utilização de um modelo dinâmico como a regressão *lagged* Poisson mostrou ser o mais efeciente para dados que apresentem uma estrutura curta de correlação temporal, quando comparado com os modelos regressão de Poisson, regressão binomial negativa e PEWMA.

A relação entre a poluição atmosférica e a mudança climática é dinâmica. As variáveis meteorológicas diretamente efetadas, de forma mais significativa pela mudança do clima são a temperatura e a humidade do ar, variáveis que servem de base para o cálculo da temperatura aparente.

Os efeitos do clima na saúde humana é um sistema aberto, composto por numerosos elementos climáticos, que atuam de forma concertada, alerta Kalkestein (1991:145-146), assim, quando os estudos possuem poucas variáveis podem dar resultados pouco fiáveis.

Na concretização deste projeto encontramos algumas dificuldades que se prendem com a falta de estudos nesta área para comparação, principlamente, nas metodologias mais sofisticadas. Estudos que comparem os métodos regressão de Poisson, regressão binomial

negativa, regressão *lagged* Poisson e PEWMA ou PAR ainda não existem nesta área de estudo. Os autores (Brandt & Williams , 1998:2000; Monogan, 2012; Forgaty & Monogan, 2012) que criaram e/ou aplicaram os modelos PAR e PEWMA e compararam com outras metodologias utilizaram dados de natureza política. Acrescendo a franca expansão da utilização dos modelos PEWMA ou PAR, que tendo sido identificados há mais de uma década ainda não se encontram instalados num *software*.

Trabalhos futuros

Uma componente importante para enriquecer este estudo e que chegou a ser ponderada, prende-se com um estudo mais completo acrescentando aos modelos variáveis socioeconómicas dos habitantes da Grande Área Metropolitana do Porto. Alguns estudos têm demonstrado que pessoas mais carenciadas são mais vulneráveis aos eventos climáticos extremos, por exemplo, por não terem ar condicionado nas habitações (WHO, 2009).

Um outro aspeto respeita à forma de recolha de dados, estes devem possuir sempre a forma mais ínfima. A base de dados apenas possui a variável mortalidade total, não permitindo comparar os grupos etários ou as causas da mortalidade, como o estudo realizado por Almeida *et.al.* (2010), que compara a mortalidade de todas as idades e por todas as causas, com a mortalidade por doenças cardivasculares e respiratórias e para os grupos etários todas as idades e maiores que 65 anos.

Deixamos mais uma sugestão dirigida às Instituições que recolhem os dados, para que os disponibilizem com maior antecipação, apenas existem estudos sobre eventos climáticos, poluição do ar e a saúde humana até ao ano 2006 e já estamos em 2012, para termos estudos de comparação mais recentes e desta forma, acompanhar a evolução actualizada dos acontecimentos.

REFERÊNCIAS

Alcoforado, M. J., Nunes, M., & Garcia, R. (1999). *A percepção da relação clima-saúde pública em Lisboa, no século XIX, através da obra de Marino Miguel Franzini.* 17(2), 31-40.

Almeida, S., Casimiro, E., & Calheiros, J. (2010). *Effects of apparent temperature on daily mortality in Lisbon and Oporto, Portugal.* Environmental Health, 9-12.

Amann, M. (2008). *Health Risks of Ozone from Long-Range Transboundary Air Pollution.* Copenhagen: WHO Regional Office Europe.

Andrade, H. (2003). *Bioclima Humano e Temperatura do Ar em Lisboa.* Tese de Doutoramento apresentada à Faculdade de Letras em Lisboa.

APA (2006). *Planning and Urban Design Standards.* New Jersey: American Planning Association. John Wiley and Sons.

Basu, R., Feng, W., & Ostro, B. (2008). *Characterizing temperature and mortality in nine California counties.* Epidemiology 2008, 19 (1), 138-145.

Bates, D. (1994). Setting the stage:critical risks. In *Environmental Health Risks and Public Policy. Decision Making in Free Societies.* Scattle: University of Wasgton Press, 6-56.

Bates, D. (1999). Introdution. In S. Holgate, J. Samet, H. Koren, & R. Maynard, *Air Pollution and Health* (pp. 1-5). London: Academic Press.

Berkson, J. (1944). *Application of the logistic funtion to bio-assay.* J. Am. Statist. Assoc. 39, 357-365.

Birch, M. W. (1963). *The detection of partial association II: the general case.* J. R. Statist. Soc. B 27, 111-124.

Bisgaard, S., & Kulahci, M. (2011). *Time Series Analysis and Forecasting by Example.* New Jersey: John Wiley & Sons.

Bliss, C. I. (1935). *The calculation of the dosage-mortality curve.* Ann. Appl. Biol. 22, 134-78.

Boor, C. D. (1978). *A Practical Guide to Splines.* New York: Springer.

Borrego, C. (1995). *Poluição Atmosférica I.* Departamento de Ambiente e Ordenamento, Universidade de Aveiro.

Braga, A., Zanobetti, A., & Schwartz, J. (2002). *The effect of weather on respiratory and cardiovascular deaths in 12 cities.* Environ Health Perspect 110, 859-863.

Brandt, P. T., & Williams , J. (1998). *Dynamic Modeling for Persistent Event Count Time Series.* Workshop in Political Theory and Policy Analysis. Indiana University, 1-37.

Brandt, P., & Williams, J. (2000). *Dynamic Modeloing for Persistent Events Count Time Series.* Workshop in Political Theory and Policy Analysis. Indiana University, 1-45.

Brandt, P., & Williams, J. (2001). *A Linear Poisson Autoregressive Model: The Poisson AR (p) Model.* Political Analysis 9 (2), 164-184.

Brockwell, P. J., & Richard A. , D. (2002). *Introduction to Time Series and Forecasting.* USA: Springer. Second Edition.

Buja, A., Hastie, T., & Tibshirani, R. (1989). *Linear Smoothers and Additive Models.* The Annals of Statistics 17 (2), 453-510.

Chapra, C., & Canale , P. (1987). *Numerical methods for engineers with personal computer applications.* EUA: McGraw-Hill International Editions.

Chatfield, C. (1996). *The Analysis on Time Series – An Introduction.* New York: Chapman & Hall. Fifth edition.

Cleveland, W. S. (1979). *Robust locally weighted regression and smoothing scatterplots.* J. Amer. Statist. Assoc. 74, 829-836.

Díaz, J., Linares, C., & Tobías, A. (2006). *Impact of extreme temperatures on daily mortality in Madrid (Spain) among the 45–64 age-group,* 342–348.

Diggle, P. (1990). *Time Series: A Biostatistical Introduction.* New York: Oxford University Press.

Dyke, G. V., & Patterson, H. D. (1952). *Analysis os factorial arrangements when the data are proportions.* Biometrics 8, 1-12.

Faraway, J. J. (2006). *Extending the Linear Model with R. Generalized Linear,Mixed Effects and Nonparametric Regression Models.* NW: Taylor & Francis Group.

Feigl, P., & Zelen, M. (1965). *Estimation of exponential survival probabilities with concomitant information.* Biometrics 21, 826-838.

Fisher, R. A. (1922). *On the mathematical foundations of theoretical statistics.* Phil. Trans. R. Soc. 222, 309-68.

Forgaty, B., & Monogan, J. (3 de April de 2012). *Modeling Time-Series Count Data: The Unique Challenges Facing Political Communication Studies.* Obtido em 1 de junho de 2012, de http://monogan.myweb.uga.edu/teaching/ts/mediaCount_v7.pdf

Friedman, J. H., & Stuetzle, W. (1981). *Projection pursuit regression.* J. Amer. Statist. Assoc. 76, 817-823.

Glasser, M. (1967). *Exponential survival with covariance.* Journal of the American Statistical Association 62, 561-568.

Gurjar, B. R., Molina, L., & Ojha, C. (2008). *Air Pollution: Health and Environmental Impacts.* USA: Taylor & Francis Group.

Hamilton, J. D. (1994). *Time Series Analysis.* USA: Princeton University Press.

Hardin, J., & Hilbe, J. (2007). *Generalized Linear Models and Extentions.* USA: Stata Press. Second Edition.

Harvey, A. C., & Fernandes, C. (1989). *Times Series Models for Count or Qualitative Observations.* Journal of Business and Economic Statistics. 7(4), 407-417.

Hastie, T., & Tibshirani, R. (1986). *Generalized Additive Models.* Statistical Science,Vol I, (3), 297-318

Hastie, T., & Tibshirani, R. (1990). *Generalized Additive Models.* New York: Chapman and Hall.

Hilbe, J. M. (2008). *Negative Binomial Regression.* UK: Cambridge University Press. Second Edition.

Hile, K. (2009). *The Handy Weather Answer Book.* USA: Visible Ink Press. Second Edition.

Hoyt, P., Fitzgerald, J., & Smith, B. (2009). *Global Experience with Problem Solving for Better Health.* New York: Springer.

Hu, W., Mengerse, K., Anthony , M., & Tong, S. (2008). *Temperature, air pollution and total mortality during summers in Sydney, 1994–2004,* 689–696.

Iñiguez, C., & et, a. (2010). *Relation between Temperature and Mortality in Thirteen Spanish Cities,* 3196-3210.

Instituto de Meteorologia, I. P. (2006). *Onda de Calor*. Obtido em 18 de agosto de 2012, de http://www.meteo.pt/pt/media/noticias/newsdetail.html?f=/pt/media/noticias/arquivo/2007/Infxclima_verao

Johnston, J., & DiNard, J. (2007). *Econometric Methods.* New York: McGraw-Hill. 4th Edition.

Kalkstein, S. L., & Valimont, K. (1986). *An Evaluation of Summer Discomfort in the United State Using a Relative Climatological Index.* Bulletin of the American Meteorological Society, vol. 67 (7), 842-848.

Kelsall, J., JM, S., Zeger, S., & Xu, J. (1997). Air pollutionand mortality in Philadelphia, 1974-1988. *Am J Epidemiol 146*, 750-762.

King, G. (1989). *Variance Specification in Event Count Models: From a Restrictive Assumptions to a Generalized Estimator.* American Journal of Political Science 33(3), 762–784.

Liang, W.-M., Liu, W.-P., & Kuo, H. (2009). Diurnal temperaturew range and emergency room admissions for cronic obstructive pulmonary disease in Taiwan. *Int J Biometeorol*, 17-23.

Maindonald, J. (2010). *Smoothing Terms in GAM Models*. Obtido em 19 de maio de 2012, de http://www.maths.anu.edu.au/~johnm/r-book/xtras/autosmooth.pdf

Maindonald, J., & Braun, J. (2007). *Data Analysis and Graphics Using R. An Example-Based Approach.* UK: Second Ediction. Cambridge University Press, 332-344.

Marques, J., & Antunes, S. (2009). A perigosidade natural da temperatura do ar em Portugal Continental: Avaliação do risco na mortalidade. *Resumos(das Comunicações apresentadas ao V Encontro Nacional e I Congresso Internacional de Riscos*, 49-61. Coimbra: Riscos - Associação Portuguesa de Riscos Prevenção e Segurança.

McCulagh, P., & Nelder , J. (1989). *Generalized Linear Models.* London: Chapman & Hall. Second Edition.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models.* London/New York. Second Edition: Chapman and Hall.

McGlade, J., & Bertollini, R. (2005). Preface by Jaqueline McGlade and Roberto Bertolini. In W. Kirch, B. Menne, & R. Bertollini, *Extreme Weather Events And Public Health Responses* (pp. XVII-XIX). New York: Springer.

Monogan, J. (11 de April de 2012). *Time Series Event Count Models.* Obtido em 20 de junho de 2012, de University of Georgia: http://monogan.myweb.uga.edu/teaching/ts/slidesTSeventCount.pdf

Monteiro, A., Carvalho, V., Oliveira, T., & Sousa, C. (2012). *Excess mortality and morbidity during the July 2006.* Int J Biometeorol.

Montero, J. C., Mirón, I. J., Criado-Álvarez, J. J., Linhares, C., & Díaz, J. (1 de Nov de 2010). Mortality From cold waves in Castile - La Mancha, Spain. *Science of the Total Environment*, 5768-5774.

Myers, R., Montgomery, D., Vining, G. G., & Robinson, T. (2010). *Generalized Linear Models with Applications inEngineering and the Sciences.* New Jersey: Wiley. Secon Edition.

Nelder, J., & Wedderburn, R. (1972). *Generalized linear models.* Journal of the Royal Statistical Society A, Vol. 135, 370–384.

Ortíz, M., Jiménez, J., Rubio , R., Odriozola, J., & Pérez, I. (1997). Mortalidad Diaria en la Comunidad de Madrid durante el periodo 1986-1991 para el grupo de edad de 45 a 64 años: su relatión con la temperatura del aire. *Rev Esp Salud Pública* , 149-160.

Peng, R., & Dominici, F. (2011). *Statistical Methods for Environmental Epidemiology with R. A Case Study in Air Pollution and Health.* USA: Springer.

Pollins, B. (1996). *Global Political Order, Economic Change, and Armed Conflict: Coevolving Systems and the Use of Force.* American Political Science Review 90(1), 103-117.

Publishing, B. E. (2011). *The Living Earth: Climate and Climate Change.* New York: Jonh P.Rafferty. First Edition.

Rothfusz, L. (1990). *The Heat Index "Equation" (or, More Than You Ever Wanted to Know About Heat Index).* Technical Attachment SR 90-23. National Weather Service Southern Region Headquarters, Fort Worth, TX.

Saez, M., Sunyer, J., Castellsagué, J., Murillo, C., & Antó, J. (1995). *Relationship between Weather Temperature and Mortality: A Time Series Analysis Approach in Barcelona.* International Journal of Epidemiolog. Vol 24, (3), 576-582.

Semenza, J., McCullough, J., Flanders, D., McGeehin, M., & Lumpkin, J. (1999). *Excess hospital admissions during the July 1995 Heat.* Am J Prev Med 16(4), 269–277.

Shumway, R. H., & Stoffer, D. (2011). *Time Series Analysis and Its Applications: With R Examples.* New York: Springer. Third edition.

Skorton, D. (2011). *Public Health in the 21st Century.* Santa Barbara - Califórnia: ABC-CLIO, LLC.

Tadano, Y. (2007). *Análise do impacto de PM10 na saúde populacional: estudo de caso em Araucária,PR.* Curitiba. 99f. Dissertação (Mestrado em Engenharia Mecânica e de Materiais) - Universidade Tecnológica de Paraná.

Turkman, M. A., & Silva, G. (2000). *VIII Congresso Anual sa Sociedade Portuguesa de Estatística. Modelos Lineares Generalizados - da teoria à prática.* Lisboa: Sociedade Portuguesa de Estatística.

Wang, X., Barnett, A., Hu, W., & Tong, S. (2009). *Temperature variation and emergency hospital admissions for stroke in Brisbane, Australia, 1996-2005.* Int J Biometeorol 53, 535-541.

Wedderburn, R. (1974). *Quasi-likelihood functions, generalized linear model and the Gauss-Newton method.* Biometrika.61(3) 439-447.

WHO (2009). *Improving public health responses to extreme weather/heat-waves EuroHEAT.* Copenhagen: World Health Organization.

Wilkinson, P., Pattenden, S., Armstrong, B., Fletcher, A., Kovats, R., Montagni, P., et al. (2004). *Vulnerability to winter mortality in eldery people in Britain: population based study.* BMJ :August 2004.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R.* USA: Chapman & Wood.

Zeger, S. (1988). *A regression model for time series of counts.* Biometrika. 75, 621-29.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). *Regression Models for Count Data in R.* Vol. 27 (8), Jul 2008.

Zippin, C., & Armitage, P. (1966). *Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter.* Biometrics 22, 665-672.